In [4]:

```
import requests # Question 2: Requests for fetching html content from website

# Make the GET request to the URL
r = requests.get('http://wikilit.referata.com/wiki/Special:Ask/-5B-5BCategory:Publications-5D-5D/-3FHas-20author%3DAuthor(s)/-3FYear/-3FPublished-20in/-3FAbstract/-3FHas-20topic%3DTopic(s)/-3FHas-20domain%3DDomain(s)/format%3D-20csv/limit%3D-20100/offset%3D0')

# Extract the content
c = r.content
```

In [5]:

```
c # Question 2: Printing the value of extracted website content
```

```
Out[5]:  b',Author(s),Year,"Published in",Abstract,Topic(s),Domain(s)\n"\'Wikipedia, the free encyclopedia\' as a r
         ole model? Lessons for open innovation from an exploratory examination of the supposedly democratic-anarch
         ic nature of Wikipedia","Gordon M\xc3\xbcller-Seitz,Guido Reger",2010,"International Journal of Technology
         Management","Accounts of open source software (OSS) development projects frequently stress their democrati
         c, sometimes even anarchic nature, in contrast to for-profit organisations. Given this observation, our re
         search evaluates qualitative data from Wikipedia, a free online encyclopaedia whose development mechanism 
         allegedly resembles that of OSS projects. Our research offers contributions to the field of open innovatio
         n research with three major findings. First, we shed light on Wikipedia as a phenomenon that has received 
         scant attention from management scholars to date. Second, we show that OSS-related motivational mechanisms
         partially apply to Wikipedia participants. Third, our exploration of Wikipedia also reveals that its organ
         isational mechanisms are often perceived as bureaucratic by contributors. This finding was unexpected sinc
         e this type of problem is often associated with for-profit organisations. Such a situation risks attenuati
         ng the motivation of contributors and sheds a critical light on the nature of Wikipedia as a role model fo
         r open innovation processes.","Contributor motivation,Policies and governance,Social order","Information s
         ystems"\n"A \'resource review\' of Wikipedia","Cormac Lawler",2006,"Counselling & Psychotherapy Researc
         h","The article offers information on Wikipedia, an online encyclopedia. The articles and definitions publ
         ished in Wikipedia can be edited. Articles usually start as a single sentence and they grow over time thro
         ugh collaborative writing and editing. A discussion page for every article is also provided for people int
         erested in or concerned with the content of that article.","Miscellaneous topics","Information system
         s"\n"A Persian web page classifier applying a combination of content-based and context-based features","Mo
         jgan Farhoodi,Alireza Yari,Maryam Mahmoudi",2009,"International Journal of Information Studies","There are
         many automatic classification methods and algorithms that have been propose for content-based or context-b
         ased features of web pages. In this paper we analyze these features and try to exploit a combination of fe
         atures to improve categorization accuracy of Persian web page classification. In this work we have suggest
         ed a linear combination of different features and adjusting the optimum weighing during application. To sh
         ow the outcome of this approach, we have conducted various experiments on a dataset consisting of all page
         s belonging to Persian Wikipedia in the field of computer. These experiments demonstrate the usefulness of
         using content-based and context-based web page features in a linear weighted combination.","Text classific
         ation","Computer science"\n"A Wikipedia literature review","Owen S. Martin",2010,ArXiv,"This paper was ori
         ginally designed as a literature review for a doctoral dissertation focusing on Wikipedia. This exposition
         gives the structure of Wikipedia and the latest trends in Wikipedia research.","Literature review",Mathema
         tics\n"A Wikipedia matching approach to contextual advertising","Alexander N. Pak,Chin-Wan Chung",2010,"Wo
         rld Wide Web","Contextual advertising is an important part of today\'s Web. It provides benefits to all pa
         rties: Web site owners and an advertising platform share the revenue, advertisers receive new customers, a
         nd Web site visitors get useful reference links. The relevance of selected ads for a Web page is essential
         for the whole system to work. Problems such as homonymy and polysemy, low intersection of keywords and con
         text mismatch can lead to the selection of irrelevant ads. Therefore, a simple keyword matching technique 
         gives a poor accuracy. In this paper, we propose a method for improving the relevance of contextual ads. W
         e propose a novel ""Wikipedia matching"" technique that uses Wikipedia articles as ""reference points"" fo
         r ads selection. We show how to combine our new method with existing solutions in order to increase the ov
         erall performance. An experimental evaluation based on a set of real ads and a set of pages from news Web 
         sites is conducted. Test results show that our proposed method performs better than existing matching stra
         tegies and using the Wikipedia matching in combination with existing approaches provides up to 50% lift in
```

```
the average precision. TREC standard measure bpref-10 also confirms the positive effect of using Wikipedia
matching for the effective ads selection.","Other information retrieval topics","Computer science"\n"A com
parison of World Wide Web resources for identifying medical information","Pamela T. Johnson,Jennifer K. Ch
en,John Eng,Martin A. Makary,Elliot K. Fishman",2008,"Academic Radiology","The objective is to compare the
utility of a search engine, Google, with other medical and non-medical, web-based resources for identifyin
g specific medical information.This institutional review board-approved case cross-over study randomly ass
igned 89 medical student volunteers to use either Google or any other web-based resource (excluding Googl
e) to research 10 advanced medical questions in a multiple choice exam. Primary outcome measures were reso
urce efficiency (inversely related to number of links used to identify the correct answer for each questio
n) and correctness (number of correct answers/total number of questions answered). For Google searches, th
e sites providing the information in question were also evaluated.The most frequently selected non-Google
resources were Yahoo (n = 531), Ask.com (n = 110), and the interactive encyclopedia Wikipedia.com (n = 7
4). Google was more efficient than all other resources (1.50 vs. 1.94 mean links, P .0001), with no signif
icant difference in correctness (97% [756/780] vs. 96% [747/780], P = .16). After a Google search, the fou
r most common categories of sites that provided the correct answer were dictionary/encyclopedia sites, med
ical websites, National Library of Medicine resources, or journal websites. Yahoo was less efficient than
Google (1.90 vs. 1.54 mean links, P .0001). However, non-Google search engines were more efficient than we
b sites (eg, Wikipedia, medical websites) and PubMed (1.87 vs. 2.54 mean links, P = .0004).Google is an ef
ficient web resource for identifying specific medical information, by guiding users to an array of medical
resources.","Health information source,Ranking and popularity",Health\n"A comparison of privacy issues in
collaborative workspaces and social networks","Martin Pek\xc3\xa1rek,Stefanie P\xc3\xb6tzsch",2009,"Identi
ty in the Information Society","With the advent of Web 2.0, numerous social software applications allow\np
eople to publish and share information on the Internet. Two of these types of\napplications \xe2\x80\x93 c
ollaborative workspaces and social network sites \xe2\x80\x93 have a number of\nfeatures in common, which
are explored to provide a basis for comparative analysis.\nThis basis is extended with a suitable definiti
on of privacy, a sociological perspective\nand an applicable adversary model in order to facilitate an inv
estigation of similarities\nand differences with regard to privacy threats. Practical examples are derived
from the\nuse of Wikipedia and Facebook. Analysis suggests that a combination of technical,\nlegal, and no
rmative solutions should be considered to counter privacy issues. A\nnumber of potential solutions that ma
y mitigate these issues are proposed.","Ethics,Policies and governance","Philosophy and ethics,Computer sc
ience,Information systems"\n"A content-driven reputation system for the Wikipedia","B. Thomas Adler,Luca d
e Alfaro",2007,"WWW \'07 Proceedings of the 16th international conference on World Wide Web","We present a
content-driven reputation system for Wikipedia authors. In our system, authors gain reputation when the ed
its they perform to Wikipedia articles are preserved by subsequent authors, and they lose reputation when
their edits are rolled back or undone in short order. Thus, author reputation is computed solely on the ba
sis of content evolution; user-to-user comments or ratings are not used. The author reputation we compute
could be used to flag new contributions from low-reputation authors, or it could be used to allow only aut
hors with high reputation to contribute to controversialor critical pages. A reputation system for the Wik
ipedia could also provide an incentive for high-quality contributions. We have implemented the proposed sy
stem, and we have used it to analyze the entire Italian and French Wikipedias, consisting of a total of 69
1, 551 pages and 5, 587, 523 revisions. Our results show that our notion of reputation has good predictive
value: changes performed by low-reputation authors have a significantly larger than average probability of
having poor quality, as judged by human observers, and of being later undone, as measured by our algorithm
```

```
s.","Reputation systems","Computer science"\n"A cultural and political economy of Web 2.0","Robert W. Geh
l",2010,"George Mason University","In this dissertation, I explore Web 2.0, an umbrella term for Web-based
software and\nservices such as blogs, wikis, social networking, and media sharing sites. This range of\nWe
b sites is complex, but is tied together by one key feature: the users of these sites and\nservices are ex
pected to produce the content included in them. That is, users write and\ncomment upon blogs, produce the
material in wikis, make connections with one another\nin social networks, and produce videos in media shar
ing sites. This has two implications.\nFirst, the increase of user-led media production has led to proclam
ations that mass media,\nhierarchy, and authority are dead, and that we are entering into a time of democr
atic\nmedia production. Second, this mode of media production relies on users to supply what\nwas traditio
nally paid labor. To illuminate this, I explore the popular media discourses\nwhich have defined Web 2.0 a
s a progressive, democratic development in media\nproduction. I consider the pleasures that users derive f
rom these sites. I then examine the\ntechnical structure of Web 2.0. Despite the arguments that present We
b 2.0 as a mass\nappropriation of the means of media production, I have found that Web 2.0 site owners\nha
ve been able to exploit users\' desires to create content and control media production.\nSite owners do th
is by deploying a dichotomous structure. In a typical Web 2.0 site, there\nis a surface, where users are f
ree to produce content and make affective connections, and\nthere is a hidden depth, where new media capit
alists convert user-generated content into\nexchange-values. Web 2.0 sites seek to hide exploitation of fr
ee user labor by limiting\naccess to this depth. This dichotomous structure is made clearer if it is compa
red to the\none Web 2.0 site where users have largely taken control of the products of their labor:\nWikip
edia. Unlike many other sites, Wikipedia allows users to see into and determine the\nlegal, technical, and
 cultural depths of that site. I conclude by pointing to the different\ncultural formations made possible
 by eliminating the barrier between surface and depth in\nWeb software architecture.","Culture and values
 of Wikipedia,Other collaboration topics,Policies and governance,Commercial aspects","Economics,Political
 science,Sociology"\n"A data-driven sketch of Wikipedia editors","Robert West,Ingmar Weber,Carlos Castill
o",2012,"WWW (Posters)","Who edits Wikipedia? We attempt to shed light on this question by using aggregate
d log data from Yahoo!\xe2\x80\x99s browser toolbar in order to analyze Wikipedians\xe2\x80\x99 editing be
havior in the context of their online lives beyond Wikipedia. We broadly characterize editors by investiga
ting how their online behavior differs from that of other users; e.g., we find that Wikipedia editors sear
ch more, read more news, play more games, and, perhaps surprisingly, are more immersed in pop culture. The
n we inspect how editors\xe2\x80\x99 general interests relate to the articles to which they contribute; e.
g., we confirm the intu-\nition that editors show more expertise in their active domains than average user
s. Our results are relevant as they illuminate novel aspects of what has become many Web users\xe2\x80\x99
 prevalent source of information and can help in recruiting new editors.","Cultural and linguistic effects
 on participation","Computer science"\n"A five-year study of on-campus Internet use by undergraduate biome
dical students","Terry Judd,Gregor Kennedy",2010,"Computers and Education","This paper reports on a five-y
ear study (2005-2009) of biomedical students\' on-campus use of the Internet. Internet usage logs were use
d to investigate students\' sessional use of key websites and technologies. The most frequented sites and
 technologies included the university\'s learning management system, Google, email and Facebook. Email was
 the primary method of electronic communication. However, its use declined over time, with a steep drop in
 use during 2006 and 2007 appearing to correspond with the rapid uptake of the social networking site Face
book. Both Google and Wikipedia gained in popularity over time while the use of other key information sour
ces, including the library and biomedical portals, remained low throughout the study. With the notable exc
eption of Facebook, most \'Web 2.0\' technologies attracted little use. The \'Net Generation\' students in
```

```
volved in this study were heavy users of generalist information retrieval tools and key online university
 services, and prefered to use externally hosted tools for online communication. These and other findings
 have important implications for the selection and provision of services by universities.","Domain-specifi
c student readership","Health,Education"\n"A framework for information quality assessment","Besiki Stvili
a,Les Gasser,Michael B. Twidale,Linda C. Smith",2007,"Journal of the American Society for Information Scie
nce and Technology","One cannot manage information quality (IQ) without first being able to measure it mea
ningfully and establishing a causal connection between the source of IQ change, the IQ problem types, the
 types of activities affected, and their implications. In this article we propose a general IQ assessment
 framework. In contrast to context-specific IQ assessment models, which usually focus on a few variables d
etermined by local needs, our framework consists of comprehensive typologies of IQ problems, related activ
ities, and a taxonomy of IQ dimensions organized in a systematic way based on sound theories and practice
s. The framework can be used as a knowledge resource and as a guide for developing IQ measurement models f
or many different settings. The framework was validated and refined by developing specific IQ measurement
 models for two large-scale collections of two large classes of information objects: Simple Dublin Core re
cords and online encyclopedia articles.","Comprehensiveness,Currency,Featured articles,Readability and sty
le","Information science,Library science"\n"A knowledge-based search engine powered by Wikipedia","David
 N. Milne,Ian H. Witten,David M. Nichols",2007,"CIKM \'07 Proceedings of the sixteenth ACM conference on C
onference on information and knowledge management","This paper describes Koru, a new search interface that
 offers effective domain-independent knowledge-based information retrieval. Koru exhibits an understanding
 of the topics of both queries and documents. This allows it to (a) expand queries automatically and (b) h
elp guide the user as they evolve their queries interactively. Its understanding is mined from the vast in
vestment of manual effort and judgment that is Wikipedia. We show how this open, constantly evolving encyc
lopedia can yield inexpensive knowledge structures that are specifically tailored to expose the topics, te
rminology and semantics of individual document collections. We conducted a detailed user study with 12 par
ticipants and 10 topics from the 2005 TREC HARD track, and found that Koru and its underlying knowledge ba
se offers significant advantages over traditional keyword search. It was capable of lending assistance to
 almost every query issued to it; making their entry more efficient, improving the relevance of the docume
nts they return, and narrowing the gap between expert and novice seekers.","Query processing","Computer sc
ience"\n"A method for category similarity calculation in wikis","Cheong-Iao Pang,Robert P. Biuk-Aghai",201
0,"Proceedings of the 6th International Symposium on Wikis and Open Collaboration (WikiSym \'10)","Wikis,
 such as Wikipedia, allow their authors to assign categories to articles in order to better organize relat
ed content. This paper presents a method to calculate similarities between categories, illustrated by a ca
lculation for the top-level categories in the Simple English version of Wikipedia.",,\n"A method for measu
ring co-authorship relationships in MediaWiki","Libby Veng-Sam Tang,Robert P. Biuk-Aghai,Simon Fong",200
8,"Proceedings of the 4th International Symposium on Wikis (WikiSym \'08)","Collaborative writing through
 wikis has become increasingly popular in recent years. When users contribute to a wiki article they impli
citly establish a co-authorship relationship. Discovering these relationships can be of value, for example
 in finding experts on a given topic. However, it is not trivial to determine the main co-authors for a gi
ven author among the potentially thousands who have contributed to a given author\'s edit history. We have
 developed a method and algorithm for calculating a co-authorship degree for a given pair of authors. We h
ave implemented this method as an extension for the MediaWiki system and demonstrate its performance which
 is satisfactory in the majority of cases. This paper also presents a method of determining an expertise g
roup for a chosen topic.",,\n"A multimethod study of information quality in wiki collaboration","Gerald C.
```

```
 Kane",2011,"ACM Transactions on Management Information Systems","In this article, the author presents the
 results of a two-phase, multimethod study of wiki-based collaboration in an attempt to better understand
 how peer-produced collaboration is done well in wiki environments. Phase 1 involves an in-depth case stud
y of the collaborative processes surrounding the development of the Wikipedia article on the 2007 Virginia
 Tech massacre. The rich data collected are used to develop an initial set of testable hypotheses of facto
rs that enhance the quality of peer-produced information in wiki environments. Phase 2 tests these theorie
s through a quantitative analysis of the collaborative features associated with 188 similar articles that
 Wikipedia considered for recognition as their best (i.e., the top 0.1%). Four collaborative features are
 examined for their effects on quality: volume of contributor activity, type of contributor activity, numb
er of anonymous contributors, and top contributor experience. Volume of contributor activity is the only f
eature that is unsupported, a particularly interesting result because previous literature connects that fa
ctor most clearly to success in wiki-based collaboration. Implications are discussed.","Antecedents of qua
lity,Featured articles,Quality improvement processes,Computational estimation of trustworthiness","Informa
tion systems"\n"A negative category based approach for Wikipedia document classification","Meenakshi Sunda
ram Murugeshan,K. Lakshmi,Saswati Mukherjee",2010,"International Journal of Knowledge Engineering and Data
 Mining","Profile based methods have been successfully used for the classification of unstructured texts.
 This paper presents a profile based method for Wikipedia XML document classification. We have used profil
es built using negative category information. Our approach exploits the structure of Wikipedia documents t
o build profiles. Two class profiles are built; one based on the whole content and the other based on the
 initial description of the Wikipedia documents. In addition, we have also explored the option of using th
e terms in the section and subsection titles. The effectiveness of cosine and fractional similarity measur
es in classifying XML documents is compared. The importance of combining two profile based classifiers is
 experimentally shown to have worked better than individual classifiers.","Text classification","Computer
 science"\n"A new year, a new Internet","Michael Castelluccio",2008,"Strategic Finance","A wiki, according
 to the guy who invented them, is the simplest online database that could possibly work. Ward Cunningham l
aunched his first wiki in 1995, and the format has been widely adopted since by academics, artists, hacker
s, and business professionals. The most famous wiki is Wikipedia, the online encyclopedia. Like other wiki
s, Wikipedia has an open editing system where the readers are the contributing editors and proofreaders. T
he readers write the articles. One of the problems with defining wikis is that the word, which actually me
ans quick"" in Hawaiian can refer to the software the community or the database. The community can be seen
 and operated as an intranet or a common workspace for collaborators. The reality is a little amorphous so
 why not go to the wiki (Wikipedia) for their take on it -- they should know.","Wikipedia as a system","Co
mputer science"\n"A practical approach to language complexity: a Wikipedia case study","Taha Yasseri,Andr
\xc3\xa1s Kornai,J\xc3\xa1nos Kert\xc3\xa9sz",2012,"PLoS ONE","In this paper we present statistical analys
is of English texts from Wikipedia. We try to address the issue of language complexity empirically by comp
aring the simple English Wikipedia (Simple) to comparable samples of the main English Wikipedia (Main). Si
mple is supposed to use a more simplified language with a limited vocabulary, and editors are explicitly r
equested to follow this guideline, yet in practice the vocabulary richness of both samples are at the same
 level. Detailed analysis of longer units (n-grams of words and part of speech tags) shows that the langua
ge of Simple is less complex than that of Main primarily due to the use of shorter sentences, as opposed t
o drastically simplified syntax or vocabulary. Comparing the two language varieties by the Gunning readabi
lity index supports this conclusion. We also report on the topical dependence of language complexity, e.g.
 that the language is more advanced in conceptual articles compared to person-based (biographical) and obj
```

```
ect-based articles. Finally, we investigate the relation between conflict and language complexity by analy
zing the content of the talk pages associated to controversial and peacefully developing articles, conclud
ing that controversy has the effect of reducing language complexity.",,\n"A request for help to improve th
e coverage of the NHS and UK healthcare issues on Wikipedia","Rod Ward",2006,"Health Information on the In
ternet","Wikipedia <http://en.wikipedia.org/>\nis an online encyclopaedia which\nanyone can edit. It has b
een suggested\nthat its coverage of the NHS and UK\nhealthcare issues is currently poor.\nTherefore, a gro
up of users have got\ntogether to create an \xe2\x80\x98NHS wikiproject\xe2\x80\x99\n<http://en.wikipedia.
org/wiki/Wikipedia:\nWikiProject_National_Health_Service>\nto try to improve this. We are trying\nto use a
 range of media to reach out\nto others with a wide range of\nknowledge and skills to ask you to\nhelp. Ex
amples of how you might do\nthis include adding information or\npictures of a hospital you know or\ndescri
bing (in an encyclopaedic way)\nan organisation you are familiar\nwith. The idea is to improve the\nqualit
y of information available to\neveryone and, as Google ranks\nWikipedia pages very highly, get\nthat infor
mation to a wide audience.","Health information source",Health\n"A semantic approach for question classifi
cation using WordNet and Wikipedia","Santosh Kumar Ray,Shailendra Singh,B.P. Joshi",2010,"Pattern Recognit
ion Letters","Question Answering Systems, unlike search engines, are providing answers to the users\' ques
tions in succinct form which requires the prior knowledge of the expectation of the user. Question classif
ication module of a Question Answering System plays a very important role in determining the expectations 
 of the user. In the literature, incorrect question classification has been cited as one of the major fact
ors for the poor performance of the Question Answering Systems and this emphasizes on the importance of qu
estion classification module designing. In this article, we have proposed a question classification method
 that exploits the powerful semantic features of the WordNet and the vast knowledge repository of the Wiki
pedia to describe informative terms explicitly. We have trained our system over a standard set of 5500 que
stions (by UIUC) and then tested it over five TREC question collections. We have compared our results with
 some standard results reported in the literature and observed a significant improvement in the accuracy o
f question classification. The question classification accuracy suggests the effectiveness of the method w
hich is promising in the field of open-domain question classification. Judging the correctness of the answ
er is an important issue in the field of question answering. In this article, we are extending question cl
assification as one of the heuristics for answer validation. We are proposing a World Wide Web based solut
ion for answer validation where answers returned by open-domain Question Answering Systems can be validate
d using online resources such as Wikipedia and Google. We have applied several heuristics for answer valid
ation task and tested them against some popular web based open-domain Question Answering Systems over a co
llection of 500 questions collected from standard sources such as TREC, the Worldbook, and the Worldfactbo
ok. The proposed method seems to be promising for automatic answer validation task.","Text classificatio
n","Computer science"\n"A systemic and cognitive view on collaborative knowledge building with wikis","Ulr
ike Cress,Joachim Kimmerle",2008,"International Journal of Computer-Supported Collaborative Learning","Wik
is provide new opportunities for learning and for collaborative knowledge\nbuilding as well as for underst
anding these processes. This article presents a theoretical\nframework for describing how learning and col
laborative knowledge building take place. In\norder to understand these processes, three aspects need to b
e considered: the social\nprocesses facilitated by a wiki, the cognitive processes of the users, and how b
oth processes\ninfluence each other mutually. For this purpose, the model presented in this article borrow
s\nfrom the systemic approach of Luhmann as well as from Piaget\xe2\x80\x99s theory of equilibration\nand
 combines these approaches. The model analyzes processes which take place in the\nsocial system of a wiki
 as well as in the cognitive systems of the users. The model also\ndescribes learning activities as proces
```

```
ses of externalization and internalization. Individual\nlearning happens through internal processes of ass
imilation and accommodation, whereas\nchanges in a wiki are due to activities of external assimilation and
 accommodation which in\nturn lead to collaborative knowledge building. This article provides empirical ex
amples for\nthese equilibration activities by analyzing Wikipedia articles. Equilibration activities are\n
described as being caused by subjectively perceived incongruities between an individuals\xe2\x80\x99\nknow
ledge and the information provided by a wiki. Incongruities of medium level cause\ncognitive conflicts whi
ch in turn activate the described processes of equilibration and\nfacilitate individual learning and colla
borative knowledge building.","Deliberative collaboration","Knowledge management,Psychology"\n"A tale of t
wo tasks: editing in the era of digital literacies","Kelly Chandler-Olcott",2009,"Journal of Adolescent & 
 Adult Literacy","This article argues that editing in the era of digital literacies is a complex, collabor
ative endeavor that requires a sophisticated awareness of audience and purpose and a knowledge of multiple
 conventions for conveying meaning and ensuring accuracy. It compares group editing of an article about th
e New York Yankees baseball team on Wikipedia, the popular online encyclopedia, to the decontextualized pr
oofreading task required of seventh graders on a state-level examination. It concludes that literacy instr
uction in schools needs to prepare students for the multiple dimensions of editing in both print and onlin
e environments, which means teaching them to negotiate meanings with others, not merely to correct surface
-feature errors.","Student contribution,Student information literacy",Education\n"A utility for estimating
 the relative contributions of wiki authors","Ofer Arazy,Eleni Stroulia",2009,"Proceedings of the 3rd Inte
rnational Conference on Weblogs and Social Media (ICWSM\xe2\x80\x9909)","Wikis were originally designed to
 hide the association between a wiki page and the authors who have produced it. However, there is evidence
 suggesting that corporate wiki users require an attribution mechanism that would automatically record (an
d present) the relative contribution of each author. In this paper we introduce an algorithm for assessing
 the contributions of wiki authors that is based on the notion of sentence ownership. The results of an em
pirical evaluation comparing the algorithm\xe2\x80\x99s output to manual evaluations reveal the type of co
ntributions captured by our algorithm. Implications for research and practice are discussed.","Collaborati
on software","Computer science"\n"Academics and Wikipedia: reframing Web 2.0 as a disruptor of traditional
 academic power-knowledge arrangements","Henk Eijkman",2010,"Campus-Wide Information Systems","Purpose - T
here is much hype about academics\' attitude to Wikipedia. This paper seeks to go beyond anecdotal evidenc
e by drawing on empirical research to ascertain how academics respond to Wikipedia and the implications th
ese responses have for the take-up of Web 2.0+. It aims to test the hypothesis that Web 2.0+, as a platfor
m built around the socially constructed nature of knowledge, is inimical to conventional power-knowledge a
rrangements in which academics are traditionally positioned as the key gatekeepers to knowledge. Design/me
thodology/approach - The research relies on quantitative and qualitative data to provide an evidence-based
 analysis of the attitudes of academics towards the student use of Wikipedia and towards Web 2.0+. These d
ata were provided via an online survey made available to a number of universities in Australia and abroad.
 As well as the statistical analysis of quantitative data, qualitative data were subjected to thematic ana
lysis using relational coding. Findings - The data by and large demonstrate that Wikipedia continues to be
 a divisive issue among academics, particularly within the soft sciences. However, Wikipedia is not as con
troversial as popular publicity would lead one to believe. Many academics use it extensively though cautio
usly themselves, and therefore tend to support a cautious approach to its use by students. However, eviden
ce supports the assertion that there is an implicit if not explicit awareness among academics that Wikiped
ia, and possibly by extension Web 2.0+, are disruptors of conventional academic power-knowledge arrangemen
ts. Practical implications - It is clear that academics respond differently to the disruptive effects that
```

```
 Web 2.0+has on the political economy of academic knowledge construction. Contrary to popular reports, res
ponses to Wikipedia are not overwhelmingly focused on resistance but encompass both cautious and creative
 acceptance. It is becoming equally clear that the increasing uptake of Web 2.0+in higher education makes
 it inevitable that academics will have to address the political consequences of this reframing of the own
ership and control of academic knowledge production. Originality/value - The paper demonstrates originalit
y and value by providing a unique, evidence-based insight into the different ways in which academics respo
nd to Wikipedia as an archetypal Web 2.0+application and by positioning Web 2.0+within the political econo
my of academic knowledge construction.","Epistemology,Knowledge source for scholars and librarians,Reader
 perceptions of credibility,Cross-domain student readership",Education\n"Accelerating networks","David M.
 D. Smith,Jukka-Pekka Onnela,Neil F. Johnson",2007,"New Journal of Physics","Evolving out-of-equilibrium n
etworks have been under intense scrutiny recently. In many real-world settings the number of links added p
er new node is not constant but depends on the time at which the node is introduced in the system. This si
mple idea gives rise to the concept of accelerating networks, for which we review an existing definition a
nd-after finding it somewhat constrictive-offer a new definition. The new definition provided here views n
etwork acceleration as a time dependent property of a given system as opposed to being a property of the s
pecific algorithm applied to grow the network. The definition also covers both unweighted and weighted net
works. As time-stamped network data becomes increasingly available, the proposed measures may be easily ap
plied to such empirical datasets. As a simple case study we apply the concepts to study the evolution of t
hree different instances of Wikipedia, namely, those in English, German, and Japanese, and find that the n
etworks undergo different acceleration regimes in their evolution.","Size of Wikipedia","Information syste
ms"\n"Access, claims and quality on the Internet - future challenges","Kim H. Veltman",2005,"Progress in I
nformatics","The vision of access to human knowledge has existed explicitly at least since the time of Ari
stotle In 1934, Otlet outlined a vision of comprehensive access to knowledge. Progress towards this vision
 entailed initial visions of hypertext, markup languages, the semantic web, Wikipedia and more recently a
 series of developments with respect to Open Source. A brief survey of these developments is provided. The
 rhetoric of the Internet insists that everything should be accessible by everyone at anytime. This poses
 obvious technical challenges and serious philosophical problems of method. If everything is accessible th
en how do we separate the chaff from the grain and how do we identify quality? Following a survey of impor
tant developments, this essay suggests five dimensions that need to be included in a future web: 1) varian
ts and multiple claims; 2) levels of certainty in making a claim; 3) levels of authority in defending a cl
aim; 4) levels of significance in assessing a claim; 5) levels of thoroughness in dealing with a claim. j
 2005 National Instiute of Informatics.","Epistemology,Reader perceptions of credibility","Information sys
tems"\n"Accuracy estimate and optimization techniques for SimRank computation","Dmitry Lizorkin,Pavel Veli
khov,Maxim Grinev,Denis Turdakov",2010,"VLDB Journal","The measure of similarity between objects is a very
 useful tool in many areas of computer science, including information retrieval. SimRank is a simple and i
ntuitive measure of this kind, based on a graph-theoretic model. SimRank is typically computed iterativel
y, in the spirit of PageRank. However, existing work on SimRank lacks accuracy estimation of iterative com
putation and has discouraging time complexity. In this paper, we present a technique to estimate the accur
acy of computing SimRank iteratively. This technique provides a way to find out the number of iterations r
equired to achieve a desired accuracy when computing SimRank. We also present optimization techniques that
 improve the computational complexity of the iterative algorithm from O(n4) in the worst case to min(O(n
l), O(n3/ log2n)), with n denoting the number of objects, and l denoting the number object-to-object relat
ionships. We also introduce a threshold sieving heuristic and its accuracy estimation that further improve
```

```
s the efficiency of the method. As a practical illustration of our techniques, we computed SimRank scores
 on a subset of English Wikipedia corpus, consisting of the complete set of articles and category link
s.","Ranking and clustering systems","Computer science"\n"Action research as a congruent methodology for u
nderstanding wikis: the case of Wikiversity","Cormac Lawler",2008,"Journal of Interactive Media in Educati
on","It is proposed that action research is an appropriate methodology for studying wikis, and is\nakin to
 research \xe2\x80\x98the wiki way\xe2\x80\x99. This proposal is contextualised within the case of Wikiver
sity, a\nproject of the Wikimedia Foundation. A framework for a participative research project is outline
d, and\nchallenges and implications of such a methodology are discussed.","Research platform",Education
\n"Adaptive indexing for content-based search in P2P systems","Aoying Zhou,Rong Zhang,Weining Qian,Quang H
ieu Vu,Tianming Hu",2008,"Data and Knowledge Engineering","One of the major challenges in Peer-to-Peer (P2
P) file sharing systems is to support content-based search. Although there have been some proposals to add
ress this challenge, they share the same weakness of using either servers or super-peers to keep global kn
owledge, which is required to identify importance of terms to avoid popular terms in query processing. As
 a result, they are not scalable and are prone to the bottleneck problem, which is caused by the high visi
ting load at the global knowledge maintainers. To that end, in this paper, we propose a novel adaptive ind
exing approach for content-based search in P2P systems, which can identify importance of terms without kee
ping global knowledge. Our method is based on an adaptive indexing structure that combines a Chord ring an
d a balanced tree. The tree is used to aggregate and classify terms adaptively, while the Chord ring is us
ed to index terms of nodes in the tree. Specifically, at each node of the tree, the system classifies term
s as either important or unimportant. Important terms, which can distinguish the node from its neighbor no
des, are indexed in the Chord ring. On the other hand, unimportant terms, which are either popular or rare
 terms, are aggregated to higher level nodes. Such classification enables the system to process queries on
 the fly without the need for global knowledge. Besides, compared to the methods that index terms separate
ly, term aggregation reduces the indexing cost significantly. Taking advantage of the tree structure, we a
lso develop an efficient search algorithm to tackle the bottleneck problem near the root. Finally, our ext
ensive experiments on both benchmark and Wikipedia datasets validated the effectiveness and efficiency of
 the proposed method.","Other information retrieval topics","Computer science"\n"Addressing gaps in knowle
dge while reading","Christopher Jordan,Carolyn Watters",2009,"Journal of the American Society for Informat
ion Science and Technology","Reading is a common everyday activity for most of us. In this article, we exa
mine the potential for using Wikipedia to fill in the gaps in one\'s own knowledge that may be encountered
 while reading. If gaps are encountered frequently while reading, then this may detract from the reader\'s
 final understanding of the given document. Our goal is to increase access to explanatory text for readers
 by retrieving a single Wikipedia article that is related to a text passage that has been highlighted. Thi
s approach differs from traditional search methods where the users formulate search queries and review lis
ts of possibly relevant results. This explicit search activity can be disruptive to reading. Our approach
 is to minimize the user interaction involved in finding related information by removing explicit query fo
rmulation and providing a single relevant result. To evaluate the feasibility of this approach, we first e
xamined the effectiveness of three contextual algorithms for retrieval. To evaluate the effectiveness for
 readers, we then developed a functional prototype that uses the text of the abstract being read as contex
t and retrieves a single relevant Wikipedia article in response to a passage the user has highlighted. We
 conducted a small user study where participants were allowed to use the prototype while reading abstract
s. The results from this initial study indicate that users found the prototype easy to use and that using
 the prototype significantly improved their stated understanding and confidence in that understanding of t
```

```
he academic abstracts they read.","Reading support","Computer science"\n"Adhocratic governance in the Inte
rnet age: a case of Wikipedia","Piotr Konieczny",2010,"Journal of Information Technology & Politics","In r
ecent years, a new realm has appeared for the study of political and sociological phenomena: the Internet.
 This article will analyze the decision-making processes of one of the largest online communities, Wikiped
ia. Founded in 2001, Wikipedia--now among the top-10 most popular sites on the Internet--has succeeded in
 attracting and organizing millions of volunteers and creating the world\'s largest encyclopedia. To date,
 however, little study has been done of Wikipedia\'s governance. There is substantial confusion about its
 decision-making structure. The organization\'s governance has been compared to many decision-making and p
olitical systems-from democracy to dictatorship, from bureaucracy to anarchy. It is the purpose of this ar
ticle to go beyond the earlier simplistic descriptions of Wikipedia\'s governance in order to advance the
 study of online governance, and of organizations more generally. As the evidence will show, while Wikiped
ia\'s governance shows elements common to many traditional governance models, it appears to be closest to
 the organizational structure known as adhocracy.","Policies and governance",Business\n"An Aesthetic for D
eliberating Online: Thinking Through \xe2\x80\x9cUniversal Pragmatics\xe2\x80\x9d and \xe2\x80\x9cDialogis
m\xe2\x80\x9d with Reference to Wikipedia","Nicholas Cimini,Jennifer Burr",2012,"The Information Society:
 An International Journal","In this article we examine contributions to Wikipedia through the prism of two
 divergent critical theorists: J\xc3\xbcrgen Habermas and Mikhail Bakhtin. We show that, in slightly dissi
milar ways, these theorists came to consider an \xe2\x80\x9caesthetic for democracy\xe2\x80\x9d (Hirschkop
 1999) or template for deliberative relationships that privileges relatively free and unconstrained dialog
ue to which every speaker has equal access and without authoritative closure. We employ Habermas\'s theory
 of \xe2\x80\x9cuniversal pragmatics\xe2\x80\x9d and Bakhtin\'s \xe2\x80\x9cdialogism\xe2\x80\x9d for anal
yses of contributions on Wikipedia for its entry on stem cells and transhumanism and show that the decisio
n to embrace either unified or pluralistic forms of deliberation is an empirical matter to be judged in so
ciohistorical context, as opposed to what normative theories insist on. We conclude by stressing the need
 to be attuned to the complexity and ambiguity of deliberative relations online.","Epistemology,Deliberati
ve collaboration","Philosophy and ethics,Rhetoric,Health,Information systems"\n"An activity theoretic mode
l for information quality change","Besiki Stvilia,Les Gasser",2008,"First Monday","To manage information q
uality (IQ) effectively, one needs to know how IQ changes over time, what causes it to change, and whether
 the changes can be predicted. In this paper we analyze the structure of IQ change in Wikipedia, an open,
 collaborative general encyclopedia. We found several patterns in Wikipedia\'s IQ process trajectories and
 linked them to article types. Drawing on the results of our analysis, we develop a general model of IQ ch
ange that can be used for reasoning about IQ dynamics in many different settings, including traditional da
tabases and information repositories.","Featured articles","Information systems"\n"An analysis of Wikipedi
a","Mohammad M. Rahman",2008,"JITTA : Journal of Information Technology Theory and Application","Wikipedia
 is defined by its founders as the ""free encyclopedia that anyone can edit."" This property we argue make
s Wikipedia a public good and hence subject to under-provision. A puzzling feature of Wikipedia however is
 its enormous size at roughly seven times that of its commercial counterparts. What is driving this growt
h? And how can we assess the reliability of this giant encyclopedia arising solely from free-editing? We m
odel contribution to Wikipedia and its reliability. We demonstrate that Wikipedia is indeed subject to fre
e-riding and offer a novel explanation for the mitigation of under-provision under such circumstances. We
 also find that the public-good feature of Wikipedia and free-riding introduce a lower-bound in the qualit
y of Wikipedia. This finding is consistent with a previous empirical study that established Wikipedia\'s s
urprisingly high level of quality. We identify Wikipedia as part of a general Internet phenomenon that we
```

```
 call the Collaborative Net and that includes features such as citizen journalism and online reviews.","An
tecedents of quality,Size of Wikipedia,Commercial aspects","Information systems"\n"An analysis of open con
tent systems","Ofer Arazy,Raymond Patterson",2005,"Proceeding of the 15th Workshop on Information Technolo
gies & Systems (WITS\xe2\x80\x9905)","Traditionally, content in organizational Knowledge Bases is created
 in a highly centralized manner to ensure quality. In Open Content Systems (OCS), on the other hand, conte
nt is generated in a distributed and decentralized manner. The most notable examples of OCS are Slashdot,
 the technology news portal, and Wikipedia, an online encyclopedia. The advantage of such systems is the s
peed in which content is accumulated, while the risk of open content systems is the lack of traditional qu
ality control mechanisms. The purpose of this paper is to examine the processes that enable an open conten
t system (OCS) to function effectively.  We conduct a survey of existing open content systems, and analyze
 the interplay between the technology underlying OCS, the user community who is responsible for content ge
neration, and the types of content managed by the OCS systems. Our analysis identifies specific settings w
here open content systems are likely to thrive.",,\n"An analysis of the delayed response to hurricane Katr
ina through the lens of knowledge management","Alton Y. K. Chua,Selcan Kaynak,Schubert S. B. Foo",2007,"Jo
urnal of the American Society for Information Science and Technology","In contrast to many recent large-sc
ale catastrophic events, such as the Turkish earthquake in 1999, the 9/11 attack in New York in 2001, the
 Bali Bombing in 2002, and the Asian Tsunami in 2004, the initial rescue effort towards Hurricane Katrina
 in the U.S. in 2005 had been sluggish. Even as Congress has promised to convene a formal inquiry into the
 response to Katrina, this article offers another perspective by analyzing the delayed response through th
e lens of knowledge management (KM). A KM framework situated in the context of disaster management is deve
loped to study three distinct but overlapping KM processes, namely, knowledge creation, knowledge transfe
r, and knowledge reuse. Drawing from a total of more than 400 documents - including local, national, and f
oreign news articles, newswires, congressional reports, and television interview transcripts, as well as I
nternet resources such as wikipedia and blogs - 14 major delay causes in Katrina are presented. The extent
 to which the delay causes were a result of the lapses in KM processes within and across the government ag
encies are discussed.","News source","Knowledge management"\n"An analysis of topical coverage of Wikipedi
a","Alexander Halavais,Derek Lackaff",2008,"Journal of Computer-Mediated Communication","Many have questio
ned the reliability and accuracy of Wikipedia. This article looks at a different but closely related one i
n the following: How broad is the coverage of Wikipedia? Differences in the interests and attention of Wik
ipedia\'s editors mean that some areas, in the traditional sciences, for example, are better covered than
 others. Two approaches to measuring this coverage are presented. The first maps the distribution of topic
s on Wikipedia to the distribution of books published. The second compares the distribution of topics in t
hree established, field-specific academic encyclopedias to the articles found in Wikipedia. Unlike the top
-down construction of traditional encyclopedias, Wikipedia\'s topical coverage is driven by the interests
 of its users, and as a result, the reliability and completeness of Wikipedia is likely to be different de
pending on the subject area of the article.",Comprehensiveness,"Information systems"\n"An axiomatic approa
ch for result diversification","Sreenivas Gollapudi,Aneesh Sharma",2009,"WWW \'09 Proceedings of the 18th
 international conference on World wide web","Understanding user intent is key to designing an effective r
anking system in a search engine. In the absence of any explicit knowledge of user intent, search engines
 want to diversify results to improve user satisfaction. In such a setting, the probability ranking princi
ple-based approach of presenting the most relevant results on top can be sub-optimal, and hence the search
 engine would like to trade-off relevance for diversity in the results. In analogy to prior work on rankin
g and clustering systems, we use the axiomatic approach to characterize and design diversification system
```

```
s. We develop a set of natural axioms that a diversification system is expected to satisfy, and show that
 no diversification function can satisfy all the axioms simultaneously. We illustrate the use of the axiom
atic framework by providing three example diversification objectives that satisfy different subsets of the
 axioms. We also uncover a rich link to the facility dispersion problem that results in algorithms for a n
umber of diversification objectives. Finally, we propose an evaluation methodology to characterize the obj
ectives and the underlying axioms. We conduct a large scale evaluation of our objectives based on two data
 sets: a data set derived from the Wikipedia disambiguation pages and a product database.","Ranking and cl
ustering systems","Computer science"\n"An empirical examination of Wikipedia\'s credibility","Thomas Chesn
ey",2006,"First Monday","Wikipedia is a free, online encyclopaedia; anyone can add content or edit existin
g content. The idea behind Wikipedia is that members of the public can add their own personal knowledge, a
nonymously if they wish. Wikipedia then evolves over time into a comprehensive knowledge base on all thing
s. Its popularity has never been questioned, although some have speculated about its authority. By its own
 admission, Wikipedia contains errors. A number of people have tested Wikipedia\'s accuracy using destruct
ive methods, that is, deliberately inserting errors. This has been criticized by Wikipedia. This short stu
dy examined Wikipedia\'s credibility by asking 258 research staff, with a response rate of 21%, to read an
 article and assess its credibility, the credibility of its author, and the credibility of Wikipedia as a
 whole. Staff were either given an article in their own expert domain or a random article. No difference w
as found between the two groups in terms of their perceived credibility of Wikipedia or of the articles\'
 authors, but a difference was found in the credibility of the articles \xe2\x80\x94 the experts found Wik
ipedia\'s articles to be more credible than the nonexperts. This suggests that the accuracy of Wikipedia i
s high. However, the results should not be seen as support for Wikipedia as a totally reliable resource a
s, according to the experts, 13% of the articles contain mistakes.",Reliability,"Information systems"\n"An
 empirical study of the effects of NLP components on geographic IR performance","Nicola Stokes,Yi Li,Alist
air Moffat,Jiawen Rong",2008,"International Journal of Geographical Information Science","Natural language
 processing (NLP) techniques, such as toponym detection and resolution, are an integral part of most geogr
aphic information retrieval (GIR) architectures. Without these components, synonym detection, ambiguity re
solution and accurate toponym expansion would not be possible. However, there are many important factors a
ffecting the success of an NLP approach to GIR, including toponym detection errors, toponym resolution err
ors and query overloading. The aim of this paper is to determine how severe these errors are in state-of-t
he-art systems, and to what extent they affect GIR performance. We show that a careful choice of weighting
 schemes in the IR engine can minimize the negative impact of these errors on GIR accuracy. We provide emp
irical evidence from the GeoCLEF 2005 and 2006 datasets to support our observations.","Geographic informat
ion retrieval","Computer science,Geography"\n"An evaluation of medical knowledge contained in Wikipedia an
d its use in the LOINC database","Jeff Friedlin,Clement J McDonald",2010,"Journal of the American Medical
 Informatics Association","The logical observation identifiers names and codes {(LOINC)} database contains
 55 000 terms consisting of more atomic components called parts. {LOINC} carries more than 18 000 distinct
 parts. It is necessary to have definitions/descriptions for each of these parts to assist users in mappin
g local laboratory codes to {LOINC.} It is believed that much of this information can be obtained from the
 internet; the first effort was with Wikipedia. This project focused on 1705 laboratory analytes (the firs
t part in the {LOINC} laboratory name). Of the 1705 parts queried, 1314 matching articles were found in Wi
kipedia. Of these, 1299 (98.9\\%) were perfect matches that exactly described the {LOINC} part, 15 (1.14\\
%) were partial matches (the description in Wikipedia was related to the {LOINC} part, but did not describ
e it fully), and 102 (7.76\\%) were mis-matches. The current release of {RELMA} and {LOINC} include Wikipe
```

```
dia descriptions of {LOINC} parts obtained as a direct result of this project.","Other information retriev
al topics","Computer science,Health"\n"An exploration on on-line mass collaboration: focusing on its motiv
ation structure","Jae Kyung Ha,Yong-Hak Kim",2009,"International Journal of Social Sciences","The Internet
 has become an indispensable part of our lives. Witnessing recent web-based mass collaboration, e.g. Wikip
edia, people are questioning whether the Internet has made fundamental changes to the society or whether i
t is merely a hyperbolic fad. It has long been assumed that collective action for a certain goal yields th
e problem of free-riding, due to its non-exclusive and non-rival characteristics. Then, thanks to recent t
echnological advances, the on-line space experienced the following changes that enabled it to produce publ
ic goods: 1) decrease in the cost of production or coordination 2) externality from networked structure 3)
 production function which integrates both self-interest and altruism. However, this research doubts the h
omogeneity of on-line mass collaboration and argues that a more sophisticated and systematical approach is
 required. The alternative that we suggest is to connect the characteristics of the goal to the motivatio
n. Despite various approaches, previous literature fails to recognize that motivation can be structurally
 restricted by the characteristic of the goal. First we draw a typology of on-line mass collaboration with
 \'the extent of expected beneficiary\' and \'the existence of externality\', and then we examine each com
bination of motivation using Benkler\'s framework. Finally, we explore and connect such typology with its
 possible dominant participating motivation.","Contributor motivation,Other collaboration topics",Sociolog
y\n"An inside view: credibility in Wikipedia from the perspective of editors","Helena Francke,Olof Sundi
n",2010,"Information Research","Introduction. The question of credibility in participatory information env
ironments, particularly Wikipedia, has been much debated. This paper investigates how editors on Swedish W
ikipedia consider credibility when they edit and read Wikipedia articles. Method. The study builds on inte
rviews with 11 editors on Swedish Wikipedia, supported by a document analysis of policies on Swedish Wikip
edia. Analysis. The interview transcripts have been coded qualitatively according to the participants\' us
e of Wikipedia and what they take into consideration in making credibility assessments. Results. The parti
cipants use Wikipedia for purposes where it is not vital that the information is correct. Their credibilit
y assessments are mainly based on authorship, verifiability, and the editing history of an article. Conclu
sions. The situations and purposes for which the editors use Wikipedia are similar to other user groups, b
ut they draw on their knowledge as members of the network of practice of wikipedians to make credibility a
ssessments, including knowledge of certain editors and of the MediaWiki architecture. Their assessments ha
ve more similarities to those used in traditional media than to assessments springing from the wisdom of c
rowds.","Contributor perceptions of credibility","Library science"\n"Analysis of community structure in Wi
kipedia","Dmitry Lizorkin,Olena Medelyan,Maria Grineva",2009,"18th int. conf. on World Wide Web (WWW)","We
 present the results of a community detection analysis of the Wikipedia graph. Distinct communities in Wik
ipedia contain semantically closely related articles. The central topic of a community can be identified u
sing PageRank. Extracted communities can be organized hierarchically similar to manually created Wikipedia
 category structure.",,\n"Analyzing and visualizing the semantic coverage of Wikipedia and its authors","T
odd Holloway,Miran Bo\xc5\xbei\xc4\x8devi\xc4\x87,Katy B\xc3\xb6rner",2007,Complexity,"This article presen
ts a novel analysis and visualization of English Wikipedia data. Our specific interest is the analysis of
 basic statistics, the identification of the semantic structure and the age of the categories in this free
 online encyclopedia, and the content coverage of its highly productive authors","Semantic relatedness","I
nformation systems"\n"Analyzing the creative editing behavior of Wikipedia editors: through dynamic social
 network analysis","Takashi Iba,Keiichi Nemoto,Bernd Peters,Peter A. Gloor",2010,"Procedia - Social and Be
havioral Sciences","This paper analyzes editing patterns of Wikipedia contributors using dynamic social ne
```

```
twork analysis. We have developed a tool that converts the edit flow among contributors into a temporal so
cial network. We are using this approach to identify the most creative Wikipedia editors among the few tho
usand contributors who make most of the edits amid the millions of active Wikipedia editors. In particula
r, we identify the key category of \xe2\x80\x9ccoolfarmers\xe2\x80\x9d, the prolific authors starting and
 building new articles of high quality. Towards this goal we analyzed the 2580 featured articles of the En
glish Wikipedia where we found two main article types: (1) articles of narrow focus created by a few subje
ct matter experts, and (2) articles about a broad topic created by thousands of interested incidental edit
ors. We then investigated the authoring process of articles about a current and controversial event. There
 we found two types of editors with different editing patterns: the mediators, trying to reconcile the dif
ferent viewpoints of editors, and the zealots, who are adding fuel to heated discussions on controversial
 topics.\n\nAs a second category of editors we look at the \xe2\x80\x9cegoboosters\xe2\x80\x9d, people who
 use Wikipedia mostly to showcase themselves. Understanding these different patterns of behavior gives imp
ortant insights about the cultural norms of online creators. In addition, identifying and policing egoboos
ters has the potential to increase the quality of Wikipedia. People best suited to enforce culture-complia
nt behavior of egoboosters through exemplary behavior and active intervention are the highly regarded cool
farmers introduced above.","Social order","Information systems"\n"Applications of semantic web methodologi
es and techniques to social networks and social websites","Sheila Kinsella,John G. Breslin,Alexandre Passa
nt,Stefan Decker",2008,"Reasoning Web","One of the most visible trends on the Web is the emergence of \xe2
\x80\x9cSocial Web\xe2\x80\x9d sites which facilitate the creation and gathering of knowledge through the
 simplification of user contributions via blogs, tagging and folksonomies, wikis, podcasts, and the deploy
ment of online social networks. The Social Web has enabled community-based knowledge acquisition with effo
rts like the Wikipedia demonstrating the \xe2\x80\x9cwisdom of the crowds\xe2\x80\x9d in creating the worl
d\xe2\x80\x99s largest online encyclopaedia. Although it is difficult to define the exact boundaries of wh
at structures or abstractions belong to the Social Web, a common property of such sites is that they facil
itate collaboration and sharing between users with low technical barriers, although usually on single site
s. As more social websites form around the connections between people and their objects of interest, and a
s these \xe2\x80\x9cobject-centred networks\xe2\x80\x9d grow bigger and more diverse, more intuitive metho
ds are needed for representing and navigating the content items in these sites: both within and across soc
ial webs ites. Also, to better enable user access to multiple sites, interoperability among social website
s is required in terms of both the content objects and the person-to-person networks expressed on each sit
e. This requires representation mechanisms to interconnect people and objects on the Social Web in an inte
roperable and extensible way. The Semantic Web provides such representation mechanisms: it can be used to
 link people and objects by representing the heterogeneous ties that bind us all to each other (either dir
ectly or indirectly). In this paper, we will describe methods that build on agreed-upon Semantic Web forma
ts to describe people, content objects, and the connections that bind them together explicitly or implicit
ly, enabling social websites to interoperate by appealing to some common semantics. We will also focus on
 how developers can use the Semantic Web to augment the ways in which they create, reuse, and link content
 on social networking sites and social websites.","Technical infrastructure","Information systems"\n"Are w
eb-based informational queries changing?","Chadwyn Tann,Mark Sanderson",2009,"Journal of the American Soci
ety for Information Science and Technology","This brief communication describes the results of a questionn
aire examining certain aspects of the Web-based information seeking practices of university students. The
 results are contrasted with past work showing that queries to Web search engines can be assigned to one o
f a series of categories: navigational, informational, and transactional. The survey results suggest that
```

```
 a large group of queries, which in the past would have been classified as informational, have become at l
east partially navigational. We contend that this change has occurred because of the rise of large Web sit
es holding particular types of information, such as Wikipedia and the Internet Movie Database.","Cross-dom
ain student readership","Computer science"\n"Arguably the greatest: sport fans and communities at work on
 Wikipedia","Meghan M. Ferriter",2009,"Sociology of Sport Journal","This article explores the socially con
structed space of Wikipedia and how the process and structure of Wikipedia enable it to act both as a vehi
cle for communication between sport fans and to subtly augment existing public narratives about sport. As
 users create article narratives, they educate fellow fans in relevant social and sport meanings. This stu
dy analyzes two aspects of Wikipedia for sports fans, application of statistical information and connectin
g athletes with other sports figures and organizations, through a discourse analysis of article content an
d the discussion pages of ten sample athletes. These pages of retired celebrity athletes provide a means f
or exploring the multidirectional production processes used by the sport fan community to celebrate record
ed events of sporting history in clearly delineated and verifiable ways, thus maintaining the sport fans\'
 community social values. Adapted from the source document.","Contributor engagement","Information system
s"\n"Art history: a guide to basic research resources","Ching-Jung Chen",2009,"Collection Building","The p
urpose of this paper is to present basic resources and practical strategies for undergraduate art history
 research. The paper is based on the author\'s experience as both an art librarian and instructor for a co
re requirement art history course. The plan detailed in this paper covers every step of the research proce
ss, from exploring the topic to citing the sources. The resources listed, which include subscription datab
ases as well as public Web sites, are deliberately limited to a manageable number. Additional topics inclu
de defining the scope of inquiry and making appropriate use of Internet resources such as Wikipedia. The p
aper provides the academic librarian with clear guidance on basic research resources in art history.","Kno
wledge source for scholars and librarians","Library science"\n"Articulations of wikiwork: uncovering value
d work in Wikipedia through barnstars","Travis Kriplean,Ivan Beschastnikh,David W. McDonald",2008,"CSCW
 \'08 Proceedings of the 2008 ACM conference on Computer supported cooperative work","Successful online co
mmunities have complex cooperative arrangements, articulations of work, and integration practices. They re
quire technical infrastructure to support a broad division of labor. Yet the research literature lacks emp
irical studies that detail which types of work are valued by participants in an online community. A conten
t analysis of Wikipedia barnstars -- personalized tokens of appreciation given to participants -- reveals
 a wide range of valued work extending far beyond simple editing to include social support, administrative
 actions, and types of articulation work. Our analysis develops a theoretical lens for understanding how w
iki software supports the creation of articulations of work. We give implications of our results for commu
nities engaged in large-scale collaborations.","Contributor engagement","Computer science"\n"Assessing the
 value of cooperation in Wikipedia","Dennis M. Wilkinson,Bernardo A. Huberman",2007,"First Monday","Since
 its inception six years ago, the online encyclopedia Wikipedia has accumulated 6.40 million articles and
 250 million edits, contributed in a predominantly undirected and haphazard fashion by 5.77 million unvett
ed volunteers. Despite the apparent lack of order, the 50 million edits by 4.8 million contributors to the
 1.5 million articles in the English-language Wikipedia follow strong certain overall regularities. We sho
w that the accretion of edits to an article is described by a simple stochastic mechanism, resulting in a
 heavy tail of highly visible articles with a large number of edits. We also demonstrate a crucial correla
tion between article quality and number of edits, which validates Wikipedia as a successful collaborative
 effort.","Featured articles,Other collaboration topics","Information systems"\n"Assigning trust to Wikipe
dia content","B. Thomas Adler,Krishnendu Chatterjee,Marco Faella,Luca de Alfaro,Ian Pye,Vishwanath Raman",
2008,"International Symposium on Wikis","The Wikipedia is a collaborative encyclopedia: anyone can contrib
ute to its articles simply by clicking on an ""edit"" button. The open nature of the Wikipedia has been ke
y to its success, but has also created a challenge: how can readers develop an informed opinion on its rel
iability? We propose a system that computes quantitative values of trust for the text in Wikipedia article
s; these trust values provide an indication of text reliability.\n\nThe system uses as input the revision
 history of each article, as well as information about the reputation of the contributing authors, as prov
ided by a reputation system. The trust of a word in an article is computed on the basis of the reputation
 of the original author of the word, as well as the reputation of all authors who edited text near the wor
d. The algorithm computes word trust values that vary smoothly across the text; the trust values can be vi
sualized using varying text-background colors. The algorithm ensures that all changes to an article\'s tex
t are reflected in the trust values, preventing surreptitious content changes.\n\nWe have implemented the
 proposed system, and we have used it to compute and display the trust of the text of thousands of article
s of the English Wikipedia. To validate our trust-computation algorithms, we show that text labeled as low
-trust has a significantly higher probability of being edited in the future than text labeled as high-trus
t.","Reputation systems","Computer science"\n"Automatic vandalism detection in Wikipedia","Martin Potthas
t,Benno Stein,Robert Gerling",2008,"European Conference on Information Retrieval","We present results of a
 new approach to detect destructive article revisions, so-called vandalism, inWikipedia. Vandalism detecti
on is a one-class classification problem, where vandalism edits are the target to be identified among all
 revisions. Interestingly, vandalism detection has not been addressed in the Information Retrieval literat
ure by now. In this paper we discuss the characteristics of vandalism as humans recognize it and develop f
eatures to render vandalism detection as a machine learning task. We compiled a large number of vandalism
 edits in a corpus, which allows for the comparison of existing and new detection approaches. Using logist
ic regression we achieve 83% precision at 77% recall with our model. Compared to the rule-based methods th
at are currently applied in Wikipedia, our approach increases the F-Measure performance by 49% while being
 faster at the same time.",Vandalism,"Computer science"\n"Automatic word sense disambiguation based on doc
ument networks","Denis Turdakov,S.D. Kuznetsov",2010,"Programming and Computer Software","In this paper, a
 survey of works on word sense disambiguation is presented, and the method used in the Texterra system [1]
 is described. The method is based on calculation of semantic relatedness of Wikipedia concepts. Compariso
n of the proposed method and the existing word sense disambiguation methods on various document collection
s is given.","Computational linguistics","Computer science"\n"Automatically assigning Wikipedia articles t
o macro-categories","Jacopo Farina,Riccardo Tasso,David Laniado",2011,"HT \'11 - Proceedings of the 22nd A
CM conference on Hypertext and hypermedia","The online encyclopedia Wikipedia offers millions of articles
 which are organized in a hierarchical category structure, created and updated by users. In this paper we
 present a technique which leverages this rich and disordered graph to assign each article to one or more
 topics. We modify an existing approach, based on the shortest paths between categories, in order to accou
nt for the direction of the hierarchy.",,\n"Automatically refining the Wikipedia infobox ontology","Fei W
u,Daniel S. Weld",2008,"Proceeding of the 17th international conference on World Wide Web","The combined e
fforts of human volunteers have recently extracted numerous facts from Wikipedia, storing them as machine-
harvestable object-attribute-value triples in Wikipedia infoboxes. Machine learning systems, such as Kyli
n, use these infoboxes as training data, accurately extracting even more semantic knowledge from natural l
anguage text. But in order to realize the full power of this information, it must be situated in a cleanly
-structured ontology. This paper introduces KOG, an autonomous system for refining Wikipedia\'s infobox-cl
ass ontology towards this end. We cast the problem of ontology refinement as a machine learning problem an
```

```
d solve it using both SVMs and a more powerful joint-inference approach expressed in Markov Logic Network
s. We present experiments demonstrating the superiority of the joint-inference approach and evaluating oth
er aspects of our system. Using these techniques, we build a rich ontology, integrating Wikipedia\'s infob
ox-class schemata with WordNet. We demonstrate how the resulting ontology may be used to enhance Wikipedia
 with improved query processing and other features.","Ontology building","Computer science"\n"Automatising
 the learning of lexical patterns: an application to the enrichment of WordNet by extracting semantic rela
tionships from Wikipedia","Maria Ruiz-Casado,Enrique Alfonseca,Pablo Castells",2007,"Data and Knowledge En
gineering","This paper describes an automatic approach to identify lexical patterns that represent semanti
c relationships between concepts in an on-line encyclopedia. Next, these patterns can be applied to extend
 existing ontologies or semantic networks with new relations. The experiments have been performed with the
 Simple English Wikipedia and WordNet 1.7. A new algorithm has been devised for automatically generalising
 the lexical patterns found in the encyclopedia entries. We have found general patterns for the hyperonym
y, hyponymy, holonymy and meronymy relations and, using them, we have extracted more than 2600 new relatio
nships that did not appear in WordNet originally. The precision of these relationships depends on the degr
ee of generality chosen for the patterns and the type of relation, being around 60-70% for the best combin
ations proposed.","Other natural language processing topics","Computer science"\n"Autonomously semantifyin
g Wikipedia","Fei Wu,Daniel S. Weld",2007,"CIKM \'07 Proceedings of the sixteenth ACM conference on Confer
ence on information and knowledge management","Berners-Lee\xe2\x80\x99s compelling vision of a Semantic We
b is hindered by a chicken-and-egg problem, which can be best solved by a bootstrapping method \xe2\x80\x9
4 creating enough structured data to motivate the development of applications. This paper argues that auto
nomously \xe2\x80\x9cSemantifying Wikipedia\xe2\x80\x9d is the best way to solve the problem. We choose Wi
kipedia as an initial data source, because it is comprehensive, not too large, high-quality, and contains
 enough manually-derived structure to bootstrap an autonomous, self-supervised process. We identify severa
l types of structures which can be automatically enhanced in Wikipedia (e.g., link structure, taxonomic da
ta, infoboxes, etc.), and we describe a prototype implementation of a self-supervised, machine learning sy
stem which realizes our vision. Preliminary experiments demonstrate the high precision of our system\xe2\x
80\x99s extracted data \xe2\x80\x94 in one case equaling that of humans.","Information extraction","Comput
er science"\n"Autopoiesis in virtual organizations","Ma\xc5\x82gorzata Pa\xc5\x84kowska",2008,"Informatica
 Economica","Virtual organizations continuously gain popularity because of the benefits created by them. G
enerally, they are defined as temporal adhocracies, project oriented, knowledge-based network organization
s. The goal of this paper is to present the hypothesis that knowledge system developed by virtual organiza
tion is an autopoietic system. The term autopoiesis"" was introduced by Maturana for self-productive syste
ms. In this paper Wikipedia is described as an example of an autopoietic system. The first part of the pap
er covers discussion on virtual organizations. Next autopoiesis\' interpretations are delivered and the va
lue of autopoiesis for governance of virtual organizations is presented. The last parts of the work compri
se short presentation of Wikipedia its principles and conclusions of Wikipedia as an autopoietic syste
m.","Wikipedia as a system",Economics\n"Avoiding tragedy in the wiki-commons","Andrew George",2007,"Virgin
ia Journal of Law and Technology","Thousands of volunteers contribute to Wikipedia, with no expectation of
 remuneration or direct credit and with the constant risk of their work being altered As a voluntary publi
c good it seems that Wikipedia ought to face a problem of noncontribution Yet Wikipedia overcomes this pro
blem, like much of the open- source movement, by locking in a core group of dedicated volunteers who are m
otivated by a desire to join and gain status within the Wikipedia community. Still, undesirable contributi
on is just as significant a risk to Wikipedia as noncontribution Bad informational inputs, including vanda
```

```
lism and anti-intellectualism, put the project at risk because Wikipedia requires a degree of credibility
 to maintain its lock-in effect. At the same time, Wikipedia is so dependent on the work of its core commu
nity that governance strategies to exclude these bad inputs must be delicately undertaken. This article ar
gues that to maximize useful participation, Wikipedia must carefully combat harmful inputs while preservin
g the zeal of its core community, as failure to do either may result in tragedy.","Contributor motivation,
Social order,Vandalism","Information systems"\n"Awarding the self in Wikipedia: identity work and the disc
losure of knowledge","Daniel Ashton",2011,"First Monday","The \xe2\x80\x98behind\xe2\x80\x93the\xe2\x80\x9
3scenes\xe2\x80\x99 discussion and edit pages of Wikipedia reveal a complex layering of debates and discus
sion between editors. Focusing on how Wikipedia \xe2\x80\x98service awards\xe2\x80\x99 can identify and di
stinguish editors, this paper explores the disclosure of knowledge as it is intimately bound up with ident
ity work. Examining contributions/edits to Wikipedia as disclosures highlights processes of identity manag
ement and work.","Quality improvement processes","Knowledge management"\n"Be nice: Wikipedia norms for sup
portive communication","Joseph M. Reagle",2010,"New Review of Hypermedia and Multimedia","Wikipedia is ack
nowledged to have been home to ""some bitter disputes"". Indeed, conflict at Wikipedia is said to be ""as
 addictive as cocaine"". Yet such observations are not cynical commentary but motivation for a collection
 of social norms. These norms speak to the intentional stance and communicative behaviors Wikipedians shou
ld adopt when interacting with one another. In the following pages I provide a survey of these norms on th
e English Wikipedia and argue that they can be characterized as supportive based on Jack Gibb\'s classic c
ommunication article ""Defensive Communication"".","Policies and governance",Communications\n"Becoming wik
ipedian: transformation of participation in a collaborative online encyclopedia","Susan L. Bryant,Andrea F
orte,Amy Bruckman",2005,"GROUP \'05 Proceedings of the 2005 international ACM SIGGROUP conference on Suppo
rting group work","Traditional activities change in surprising ways when computer-mediated communication b
ecomes a component of the activity system. In this descriptive study, we leverage two perspectives on soci
al activity to understand the experiences of individuals who became active collaborators in Wikipedia, a p
rolific, cooperatively-authored online encyclopedia. Legitimate peripheral participation provides a lens f
or understanding participation in a community as an adaptable process that evolves over time. We use ideas
 from activity theory as a framework to describe our results. Finally, we describe how activity on the Wik
ipedia stands in striking contrast to traditional publishing and suggests a new paradigm for collaborative
 systems.","Contributor engagement","Information systems"\n"Beyond Google: how do students conduct academi
c research?","Alison J. Head",2007,"First Monday","This paper reports findings from an exploratory study a
bout how students majoring in humanities and social sciences use the Internet and library resources for re
search. Using student discussion groups, content analysis, and a student survey, our results suggest stude
nts may not be as reliant on public Internet sites as previous research has reported. Instead, students in
 our study used a hybrid approach for conducting course\xe2\x80\x93related research. A majority of student
s leveraged both online and offline sources to overcome challenges with finding, selecting, and evaluating
 resources and gauging professors\xe2\x80\x99 expectations for quality research.","Cross-domain student re
adership","Education,Library science"\n"Beyond the legacy of the enlightenment? Online encyclopaedias as d
igital heterotopias","Jutta Haider,Olof Sundin",2010,"First Monday","This article explores how we can unde
rstand contemporary participatory online encyclopaedic expressions, particularly Wikipedia, in their tradi
tional role as continuation of the Enlightenment ideal, as well as in the distinctly different space of th
e Internet. Firstly we position these encyclopaedias in a historical tradition. Secondly, we assign them a
 place in contemporary digital networks which marks them out as sites in which Enlightenment ideals of uni
versal knowledge take on a new shape. We argue that the Foucauldian concept of heterotopia, that is specia
```

```
l spaces which exist within society, transferred online, can serve to understand Wikipedia and similar par
ticipatory online encyclopaedias in their role as unique spaces for the construction of knowledge, memory
 and culture in late modern society.","Encyclopedias,Epistemology","Philosophy and ethics,Information scie
nce,Library science"\n"Beyond vandalism: Wikipedia trolls","Pnina Shachaf,Noriko Hara",2010,"Journal of In
formation Science","Research on trolls is scarce, but their activities challenge online communities; one o
f the main challenges of the Wikipedia community is to fight against vandalism and trolls. This study iden
tifies Wikipedia trolls behaviours and motivations, and compares and contrasts hackers with trolls; it ext
ends our knowledge about this type of vandalism and concludes that Wikipedia trolls are one type of hacke
r. This study reports that boredom, attention seeking, and revenge motivate trolls; they regard Wikipedia
 as an entertainment venue, and find pleasure from causing damage to the community and other people. Findi
ngs also suggest that trolls behaviours are characterized as repetitive, intentional, and harmful actions
 that are undertaken in isolation and under hidden virtual identities, involving violations of Wikipedia p
olicies, and consisting of destructive participation in the community.","Contributor motivation,Vandalis
m","Information systems"\n"BinRank: scaling dynamic authority-based search using materialized subgraph
s","Heasoo Hwang,Andrey Balmin,Berthold Reinwald,Erik Nijkamp",2010,"IEEE Transactions on Knowledge and Da
ta Engineering","Dynamic authority-based keyword search algorithms, such as ObjectRank and personalized Pa
geRank, leverage semantic link information to provide high quality, high recall search in databases, and t
he Web. Conceptually, these algorithms require a query-time PageRank-style iterative computation over the
 full graph. This computation is too expensive for large graphs, and not feasible at query time. Alternati
vely, building an index of precomputed results for some or all keywords involves very expensive preprocess
ing. We introduce BinRank, a system that approximates ObjectRank results by utilizing a hybrid approach in
spired by materialized views in traditional query processing. We materialize a number of relatively small
 subsets of the data graph in such a way that any keyword query can be answered by running ObjectRank on o
nly one of the subgraphs. BinRank generates the subgraphs by partitioning all the terms in the corpus base
d on their co-occurrence, executing ObjectRank for each partition using the terms to generate a set of ran
dom walk starting points, and keeping only those objects that receive non-negligible scores. The intuition
 is that a subgraph that contains all objects and links relevant to a set of related terms should have all
 the information needed to rank objects with respect to one of these terms. We demonstrate that BinRank ca
n achieve subsecond query execution time on the English Wikipedia data set, while producing high-quality s
earch results that closely approximate the results of ObjectRank on the original graph. The Wikipedia link
 graph contains about 108 edges, which is at least two orders of magnitude larger than what prior state of
 the art dynamic authority-based search systems have been able to demonstrate. Our experimental evaluation
 investigates the trade-off between query execution time, quality of the results, and storage requirements
 of BinRank.","Query processing,Ranking and clustering systems","Computer science"\n"Biographical social n
etworks on Wikipedia - a cross-cultural study of links that made history","Pablo Arag\xc3\xb3n,Andreas Kal
tenbrunner,David Laniado,Yana Volkovich",2012,WikiSym,"It is arguable whether history is made by great men
 and women or vice versa, but undoubtably social connections shape history. Analysing Wikipedia, a global
 collective memory place, we aim to understand how social links are recorded across cultures. Starting wit
h the set of biographies in the English Wikipedia we focus on the networks of links between these biograph
ical articles on the 15 largest language Wikipedias. We detect the most central characters in these networ
ks and point out culture-related peculiarities. Furthermore, we reveal remarkable similarities between dis
tinct groups of language Wikipedias and highlight the shared knowledge about connections between persons a
cross cultures.","Other content topics","Computer science,Information systems"\n"Breaking the knowledge ac
```

```
quisition bottleneck through conversational knowledge management","Christian Wagner",2005,"Information Res
ources Management Journal","Much of today\xe2\x80\x99s organizational knowledge still exists outside of fo
rmal information repositories and often only in people\xe2\x80\x99s heads. While organizations are eager t
o capture this knowledge, existing acquisition methods are not up to the task. Neither traditional artific
ial intelligence-based approaches nor more recent, less-structured knowledge management techniques have ov
ercome the knowledge acquisition challenges. This article investigates knowledge acquisition bottlenecks a
nd proposes the use of collaborative, conversational knowledge management to remove them. The article demo
nstrates the opportunity for more effective knowledge acquisition through the application of the principle
s of Bazaar style, open-source development. The article introduces wikis as software that enables this typ
e of knowledge acquisition. It empirically analyzes the Wikipedia to produce evidence for the feasibility
 and effectiveness of the proposed approach.","Other content topics,Quality improvement processes,Other pa
rticipation outcomes","Knowledge management"\n"Bridging domains using world wide knowledge for transfer le
arning","Evan Wei Xiang,Bin Cao,Derek Hao Hu,Qiang Yang",2010,"IEEE Transactions on Knowledge and Data Eng
ineering","A major problem of classification learning is the lack of ground-truth labeled data. It is usua
lly expensive to label new data instances for training a model. To solve this problem, domain adaptation i
n transfer learning has been proposed to classify target domain data by using some other source domain dat
a, even when the data may have different distributions. However, domain adaptation may not work well when
 the differences between the source and target domains are large. In this paper, we design a novel transfe
r learning approach, called BIG (Bridging Information Gap), to effectively extract useful knowledge in a w
orldwide knowledge base, which is then used to link the source and target domains for improving the classi
fication performance. BIG works when the source and target domains share the same feature space but differ
ent underlying data distributions. Using the auxiliary source data, we can extract a bridge that allows cr
oss-domain text classification problems to be solved using standard semisupervised learning algorithms. A
 major contribution of our work is that with BIG, a large amount of worldwide knowledge can be easily adap
ted and used for learning in the target domain. We conduct experiments on several real-world cross-domain
 text classification tasks and demonstrate that our proposed approach can outperform several existing doma
in adaptation approaches significantly.","Text classification","Computer science"\n"Building semantic kern
els for text classification using Wikipedia","Pu Wang,Carlotta Domeniconi",2008,"International Conference
 on Knowledge Discovery and Data Mining","Document classification presents difficult challenges due to the
 sparsity and the high dimensionality of text data, and to the complex semantics of the natural language.
 The traditional document representation is a word-based vector (Bag of Words, or BOW), where each dimensi
on is associated with a term of the dictionary containing all the words that appear in the corpus. Althoug
h simple and commonly used, this representation has several limitations. It is essential to embed semantic
 information and conceptual patterns in order to enhance the prediction capabilities of classification alg
orithms. In this paper, we overcome the shortages of the BOW approach by embedding background knowledge de
rived from Wikipedia into a semantic kernel, which is then used to enrich the representation of documents.
 Our empirical evaluation with real data sets demonstrates that our approach successfully achieves improve
d classification accuracy with respect to the BOW technique, and to other recently developed methods.","Te
xt classification","Computer science"\n"Can history be open source? Wikipedia and the future of the pas
t","Roy Rosenzweig",2006,"Journal of American History","The article presents information on Wikipedia, an
 online encyclopedia that contains articles about history. Wikipedia allows Internet users to freely read
 and use articles, thus, making it the most significant application of the principles of the free and open
-source software movement to the world of cultural production. Astonishingly, Wikipedia has become widely
```

```
 read and cited, with more than a million people a day visiting the site. The article also offers informat
ion on other Web-based encyclopedias that were developed before Wikipedia.","Antecedents of quality,Compre
hensiveness,Reliability,Research platform",History\n"Categorising social tags to improve folksonomy-based 
 recommendations","Iv\xc3\xa1n Cantador,Ioannis Konstas,Joemon M. Jose",2011,"Journal of Web Semantics","I
n social tagging systems, users have different purposes when they annotate items. Tags not only depict the
 content of the annotated items, for example by listing the objects that appear in a photo, or express con
textual information about the items, for example by providing the location or the time in which a photo wa
s taken, but also describe subjective qualities and opinions about the items, or can be related to organis
ational aspects, such as self-references and personal tasks.\n\nCurrent folksonomy-based search and recomm
endation models exploit the social tag space as a whole to retrieve those items relevant to a tag-based qu
ery or user profile, and do not take into consideration the purposes of tags. We hypothesise that a signif
icant percentage of tags are noisy for content retrieval, and believe that the distinction of the personal
 intentions underlying the tags may be beneficial to improve the accuracy of search and recommendation pro
cesses.\n\nWe present a mechanism to automatically filter and classify raw tags in a set of purpose-orient
ed categories. Our approach finds the underlying meanings (concepts) of the tags, mapping them to semantic
 entities belonging to external knowledge bases, namely WordNet and Wikipedia, through the exploitation of
 ontologies created within the W3C Linking Open Data initiative. The obtained concepts are then transforme
d into semantic classes that can be uniquely assigned to content- and context-based categories. The identi
fication of subjective and organisational tags is based on natural language processing heuristics.\n\nWe c
ollected a representative dataset from Flickr social tagging system, and conducted an empirical study to c
ategorise real tagging data, and evaluate whether the resultant tags categories really benefit a recommend
ation model using the Random Walk with Restarts method. The results show that content- and context-based t
ags are considered superior to subjective and organisational tags, achieving equivalent performance to usi
ng the whole tag space.","Ontology building","Computer science"\n"Characterization and prediction of Wikip
edia edit wars","R\xc3\xb3bert Sumi,Taha Yasseri,Andr\xc3\xa1s Rung,Andr\xc3\xa1s Kornai,J\xc3\xa1nos Kert
\xc3\xa9sz",2011,"In: Proceedings of the ACM WebSci\'11, June 14-17 2011, Koblenz, Germany.","We present a
 new, ecient method for automatically de- tecting conflict cases and test it on ve dierent language Wikipe
dias. We discuss how the number of edits, reverts, the length of discussions deviate in such pages from th
ose following the general workflow.",,\n"Characterizing and modeling the dynamics of online popularity","J
acob Ratkiewicz,Santo Fortunato,Alessandro Flammini,Filippo Menczer,Alessandro Vespignani",2010,"Physical 
 Review Letters","Online popularity has an enormous impact on opinions, culture, policy, and profits. We p
rovide a quantitative, large scale, temporal analysis of the dynamics of online content popularity in two 
 massive model systems: the Wikipedia and an entire country\'s Web space. We find that the dynamics of pop
ularity are characterized by bursts, displaying characteristic features of critical systems such as fat-ta
iled distributions of magnitude and interevent time. We propose a minimal model combining the classic pref
erential popularity increase mechanism with the occurrence of random popularity shifts due to exogenous fa
ctors. The model recovers the critical features observed in the empirical analysis of the systems analyzed
 here, highlighting the key factors needed in the description of popularity dynamics.","Ranking and popula
rity","Computer science"\n"Chemical information media in the chemistry lecture hall: a comparative assessm
ent of two online encyclopedias","Lukas Korosec,Peter Andreas Limacher,Hans Peter L\xc3\xbcthi,Martin Paul
 Br\xc3\xa4ndle",2010,CHIMIA,"The chemistry encyclopedia Roempp Online and the German universal encycloped
ia Wikipedia were assessed by first-year university students on the basis of a set of 30 articles about ch
emical thermodynamics. Criteria with regard to both content and form were applied in the comparison; 619 r
```

```
atings (48\\% participation rate) were returned. While both encyclopedias obtained very good marks and per
formed nearly equally with regard to their accuracy, the average overall mark for Wikipedia was better tha
n for Roempp Online, which obtained lower marks with regard to completeness and length. Analysis of the re
sults and participants\' comments shows that students attach importance to completeness, length and compre
hensibility rather than accuracy, and also attribute less value to the availability of sources which valid
ate an encyclopedia article. Both encyclopedias can be promoted as a starting reference to access a topic
 in chemistry. However, it is recommended that instructors should insist that students do not rely solely
 on encyclopedia texts, but use and cite primary literature in their reports.","Comprehensiveness,Readabil
ity and style,Reader perceptions of credibility,Domain-specific student readership","Chemistry,Educatio
n"\n"Circadian patterns of Wikipedia editorial activity: A demographic analysis","Taha Yasseri,R\xc3\xb3be
rt Sumi,J\xc3\xa1nos Kert\xc3\xa9sz",2012,"PLoS ONE","Wikipedia (WP) as a collaborative, dynamical system
 of humans is an appropriate subject of social studies. Each single action of the members of this society,
 i.e., editors, is well recorded and accessible. Using the cumulative data of 34 Wikipedias in different l
anguages, we try to characterize and find the universalities and differences in temporal activity patterns
 of editors. Based on this data, we estimate the geographical distribution of editors for each WP in the g
lobe. Furthermore we also clarify the differences among different groups of WPs, which originate in the va
riance of cultural and social features of the communities of editors.",,\n"Classifying tags using open con
tent resources","Simon Overell,B\xc3\xb6rkur Sigurbj\xc3\xb6rnsson,Roelof Van Zwol",2009,"WSDM \'09 Procee
dings of the Second ACM International Conference on Web Search and Data Mining","Tagging has emerged as a
 popular means to annotate on-line objects such as bookmarks, photos and videos. Tags vary in semantic mea
ning and can describe different aspects of a media object. Tags describe the content of the media as well
 as locations, dates, people and other associated meta-data. Being able to automatically classify tags int
o semantic categories allows us to understand better the way users annotate media objects and to build too
ls for viewing and browsing the media objects. In this paper we present a generic method for classifying t
ags using third party open content resources, such as Wikipedia and the Open Directory. Our method uses st
ructural patterns that can be extracted from resource meta-data. We describe the implementation of our met
hod on Wikipedia using WordNet categories as our classification schema and ground truth. Two structural pa
tterns found in Wikipedia are used for training and classification: categories and templates. We apply our
 system to classifying Flickr tags. Compared to a WordNet baseline our method increases the coverage of th
e Flickr vocabulary by 115%. We can classify many important entities that are not covered by WordNet, such
 as, London Eye, Big Island, Ronaldinho, geo-caching and wii.","Text classification","Computer scienc
e"\n"Clustering of scientific citations in Wikipedia","Finn \xc3\x85rup Nielsen",2008,Wikimania,"The insta
nces of templates in Wikipedia form an interesting data set of structured information. Here I focus on the
 cite journal template that is primarily used for citation to articles in scientific journals. These citat
ions can be extracted and analyzed: Non-negative matrix factorization is performed on a (article x journa
l) matrix resulting in a soft clustering of Wikipedia articles and scientific journals, each cluster more
 or less representing a scientific topic.","Reliability,Ranking and clustering systems","Computer science,
Information science"\n"Clustering short texts using Wikipedia","Somnath Banerjee,Krishnan Ramanathan,Ajay
 Gupta",2007,"SIGIR \'07 Proceedings of the 30th annual international ACM SIGIR conference on Research and
 development in information retrieval","Subscribers to the popular news or blog feeds (RSS/Atom) often fac
e the problem of information overload as these feed sources usually deliver large number of items periodic
ally. One solution to this problem could be clustering similar items in the feed reader to make the inform
ation more manageable for a user. Clustering items at the feed reader end is a challenging task as usually
```

```
 only a small part of the actual article is received through the feed. In this paper, we propose a method
 of improving the accuracy of clustering short texts by enriching their representation with additional fea
tures from Wikipedia. Empirical results indicate that this enriched representation of text items can subst
antially improve the clustering accuracy when compared to the conventional bag of words representatio
n.","Ranking and clustering systems","Computer science"\n"Co-authorship 2.0: patterns of collaboration in
 Wikipedia","David Laniado,Riccardo Tasso",2011,"HT \'11 - Proceedings of the 22nd ACM conference on Hyper
text and hypermedia","The study of collaboration patterns in wikis can help shed light on the process of c
ontent creation by online communities. To turn a wiki\'s revision history into a collaboration network, we
 propose an algorithm that identifies as authors of a page the users who provided the most of its relevant
 content, measured in terms of quantity and of acceptance by the community. The scalability of this approa
ch allows us to study the English Wikipedia community as a co-authorship network. We find evidence of the
 presence of a nucleus of very active contributors, who seem to spread over the whole wiki, and to interac
t preferentially with inexperienced users. The fundamental role played by this elite is witnessed by the g
rowing centrality of sociometric stars in the network. Isolating the community active around a category, i
t is possible to study its specific dynamics and most influential authors.","Other collaboration topics,Pa
rticipation trends,Reputation systems",\n"Codifying collaborative knowledge: using Wikipedia as a basis fo
r automated ontology learning","Tao Guo,David G. Schwartz,Frada Burstein,Henry Linger",2009,"Knowledge Man
agement Research & Practice","In the context of knowledge management, ontology construction can be conside
red as a part of capturing of the body of knowledge of a particular problem domain. Traditionally, ontolog
y construction assumes a tedious codification of the domain experts knowledge. In this paper, we describe
 a new approach to ontology engineering that has the potential of bridging the dichotomy between codificat
ion and collaboration turning to Web 2.0 technology. We propose to shift the primary source of ontology kn
owledge from the expert to socially emergent bodies of knowledge such as Wikipedia. Using Wikipedia as an
 example, we demonstrate how core terms and relationships of a domain ontology can be distilled from this
 socially constructed source. As an illustration, we describe how our approach achieved over 90\\% concept
ual coverage compared with Gold standard hand-crafted ontologies, such as Cyc. What emerges is not a folks
onomy, but rather a formal ontology that has nonetheless found its roots in social knowledge.","Ontology b
uilding","Information science,Knowledge management"\n"Collaboration in context: comparing article evolutio
n among subject disciplines in Wikipedia","Katherine Ehmann,Andrew Large,Jamshid Beheshti",2008,"First Mon
day","This exploratory study examines the relationships between article and talk page contributions and th
eir effect on article quality in Wikipedia. The sample consisted of three articles each from the hard scie
nces, soft sciences, and humanities, whose talk page and article edit histories were observed over a five-
month period and coded for contribution types. Richness and neutrality criteria were then used to assess a
rticle quality and results were compared within and among subject disciplines. This study reveals variabil
ity in article quality across subject disciplines and a relationship between talk page discussion and arti
cle editing activity. Overall, results indicate the initial article creator\'s critical role in providing
 a framework for future editing as well as a remarkable stability in article content over time.","Antecede
nts of quality,Readability and style,Reliability,Other collaboration topics","Information systems"\n"Colla
borative authoring on the web: a genre analysis of online encyclopedias","William Emigh,Susan C. Herring",
2005,"HICSS \'05 Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on Syst
em Sciences (HICSS\'05) - Track 4 - Volume 04","This paper presents the results of a genre analysis of two
 web-based collaborative authoring environments, Wikipedia and Everything2, both of which are intended as
 repositories of encyclopedic knowledge and are open to contributions from the public. Using corpus lingui
```

```
stic methods and factor analysis of word counts for features of formality and informality, we show that th
e greater the degree of post-production editorial control afforded by the system, the more formal and stan
dardized the language of the collaboratively-authored documents becomes, analogous to that found in tradit
ional print encyclopedias. Paradoxically, users who faithfully appropriate such systems create homogeneous
 entries, at odds with the goal of open-access authoring environments to create diverse content. The findi
ngs shed light on how users, acting through mechanisms provided by the system, can shape (or not) features
 of content in particular ways. We conclude by identifying sub-genres of web-based collaborative authoring
 environments based on their technical affordances.","Readability and style","Information systems"\n"Colle
ctivism vs. individualism in a wiki world: librarians respond to Jaron Lanier\'s essay \'Digital Maoism: t
he hazards of the new online collectivism\'","Markel Tumlin,Steven R. Harris,Heidi Buchanan,Krista Schmid
t,Kay Johnson",2007,"Serials Review","Jaron Lanier\'s essay {Digital} Maoism: The Hazards of the New Onlin
e Collectivism"" is a self-described rant of the dangers of the hive mentality in suppressing individual h
uman intelligence as demonstrated in online resources such as Wikipedia and {MySpace.} He sees merit in co
llective decision-making and problem-solving if evaluation is uncontroversial but argues that individuals
 are essential in providing judgment taste and user experiences in many situations. Lanier\'s essay appear
ed in the online progressive publication Edge and received responses from a variety of technologists acade
mics and writers. In this {""Balance} Point"" column four academic librarians provide a library public ser
vices viewpoint in responding to Lanier\'s essay.""","Other collaboration topics","Library science"\n"Comm
ons-based peer production and virtue","Yochai Benkler,Helen Nissenbaum",2006,"The Journal of Political Phi
losophy","COMMONS-BASED peer production is a socio-economic system of production that is emerging in the d
igitally networked environment. Facilitated by the technical infrastructure of the Internet, the hallmark
 of this socio-technical system is collaboration among large groups of individuals, sometimes in the order
 of tens or even hundreds of thousands, who cooperate effectively to provide information, knowledge or cul
tural goods without relying on either market pricing or managerial hierarchies to coordinate their common
 enterprise.1 While there are many practical reasons to try to understand a novel system of production tha
t has produced some of the finest software, the fastest supercomputer and some of the best web-based direc
tories and news sites, here we focus on the ethical, rather than the functional dimension. What does it me
an in ethical terms that many individuals can find themselves cooperating productively with strangers and
 acquaintances on a scope never before seen? How might it affect, or at least enable, human action and aff
ection, and how would these effects or possibilities affect our capacities to be virtuous human beings? We
 suggest that the emergence of peer production offers an opportunity for more people to engage in practice
s that permit them to exhibit and experience virtuous behavior. We posit: (a) that a society that provides
 opportunities for virtuous behavior is one that is more conducive to virtuous individuals; and (b) that t
he practice of effective virtuous behavior may lead to more people adopting virtues as their own, or as at
tributes of what they see as their self-definition. The central thesis of this paper is that socio-technic
al systems of commons-based peer production offer not only a remarkable medium of production for various k
inds of information goods but serve as a context for positive character formation. Exploring and substanti
ating these claims will be our quest, but we begin with a brief tour through this strange and exciting new
 landscape of commons-based peer production and conclude with recommendations for public policy.","Deliber
ative collaboration","Information systems"\n"Community building around encyclopaedic knowledge","Josef Kol
bitsch,Hermann Maurer",2004,"Journal of Computing and Information Technology","This paper gives a brief ov
erview of current technologies in systems handling encyclopaedic knowledge. Since most of the electronic e
ncyclopaedias currently available are rather static and inflexible, greatly enhanced func- tionality is in
```

```
troduced that enables users to work more effectively and collaboratively. Users have the ability, for inst
ance, to add annotations to every kind of object and can have private and shared workspaces. The technique
s described employ user profiles in order to adapt to users and involve statistical analysis to improve se
arch results. Moreover, a tracking and navigation mechanism based on trails is presented. The second part
 of the paper details community building around encyclopaedic knowledge with the aim to involve \xc3\xa2\x
e2\x82\xac\xc5\x93plain\xc3\xa2\xe2\x82\xac? users and experts in environments with largely editorial cont
ent. The foundations for building a user community are specified along with significant facets such as ret
aining the high quality of content, rating mech- anisms and social aspects. A system that implements large
 portions of the community-related concepts in a heterogeneous environment of several largely indepen- den
t data sources is proposed. Apart from online and {DVD-based} encyclopaedias, potential application areas
 are {e-Learning}, corporate documentation and knowledge management systems.","Encyclopedias,Technical inf
rastructure,Community building","Information systems"\n"Community, consensus, coercion, control: cs*w or h
ow policy mediates mass participation","Travis Kriplean,Ivan Beschastnikh,David W. McDonald,Scott A. Golde
r",2007,"GROUP \'07 Proceedings of the 2007 international ACM conference on Supporting group work","When l
arge groups cooperate, issues of conflict and control surface because of differences in perspective. Manag
ing such diverse views is a persistent problem in cooperative group work. The Wikipedian community has res
ponded with an evolving body of policies that provide shared principles, processes, and strategies for col
laboration. We employ a grounded approach to study a sample of active talk pages and examine how policies
 are employed as contributors work towards consensus. Although policies help build a stronger community, w
e find that ambiguities in policies give rise to power plays. This lens demonstrates that support for mass
 collaboration must take into account policy and power.","Policies and governance","Information system
s"\n"Community-based knowledge production: antecedents of product quality in Wikipedia","Ofer Arazy,Oded N
ov",2008,"6th Annual International Open and User Innovation Workshop","Recent years have seen the emergenc
e of a new community-based model for the production of knowledge-based goods. The primary examples of this
 model are open source software and open content systems, such as Wikipedia. While the community-based mod
el has been studied extensively, relatively little is known about the factors driving the quality of commu
nity-produced goods. The objective of this paper is to develop a theoretical model of product quality dete
rminants for the community-based model, focusing on the task-related conflict that is associated with coll
ective production process. We report on an empirical study of Wikipedia that supports the model\xe2\x80\x9
9s hypotheses, and demonstrates that task conflict drives product quality, and is affected by group divers
ity and group members\xe2\x80\x99 commitment to the community. We conclude by discussing implications for
 the study of community-based knowledge production.",,\n"Comparing featured article groups and revision pa
tterns correlations in Wikipedia","Giacomo Poderi",2009,"First Monday","Collaboratively written by thousan
ds of people, Wikipedia produces entries which are consistent with criteria agreed by Wikipedians and of h
igh quality. This article focuses on Wikipedia\'s featured articles and shows that not every contribution
 can be considered as being of equal quality. Two groups of articles are analysed by focusing on the edits
 distribution and the main editors\' contribution. The research shows how these aspects of the revision pa
tterns can change dependent upon the category to which the articles belong.","Featured articles,Other coll
aboration topics","Information systems"\n"Comparing methods for single paragraph similarity analysis","Ben
jamin Stone,Simon Dennis,Peter J. Kwantes",2010,"Topics in Cognitive Science","Abstract The focus of this
 paper is two-fold. First, similarities generated from six semantic models were compared to human ratings
 of paragraph similarity on two datasets\xc3\xa2\xe2\x82\xac\xe2\x80\x9d23 World Entertainment News Networ
k paragraphs and 50 {ABC} newswire paragraphs. Contrary to findings on smaller textual units such as word
```

```
 associations {(Griffiths}, Tenenbaum, \\& Steyvers, 2007), our results suggest that when single paragraph
s are compared, simple nonreductive models (word overlap and vector space) can provide better similarity e
stimates than more complex models {(LSA}, Topic Model, {SpNMF}, and {CSM).} Second, various methods of cor
pus creation were explored to facilitate the semantic models\xc3\xa2\xe2\x82\xac\xe2\x84\xa2 similarity es
timates. Removing numeric and single characters, and also truncating document length improved performance.
 Automated construction of smaller Wikipedia-based corpora proved to be very effective, even improving upo
n the performance of corpora that had been chosen for the domain. Model performance was further improved b
y augmenting corpora with dataset paragraphs.","Other natural language processing topics",Psychology\n"Com
parison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles","Luc
y Holman Rector",2008,"Reference Services Review","This paper seeks to provide reference librarians and fa
culty with evidence regarding the comprehensiveness and accuracy of Wikipedia articles compared with respe
cted reference resources. This content analysis evaluated nine Wikipedia articles against comparable artic
les in Encyclopaedia Britannica, The Dictionary of American History and American National Biography Online
 in order to compare Wikipedia\'s comprehensiveness and accuracy. The researcher used a modification of a
 stratified random sampling and a purposive sampling to identify a variety of historical entries and compa
red each text in terms of depth, accuracy, and detail. The study did reveal inaccuracies in eight of the n
ine entries and exposed major flaws in at least two of the nine Wikipedia articles. Overall, Wikipedia\'s
 accuracy rate was 80 percent compared with 95-96 percent accuracy within the other sources. This study do
es support the claim that Wikipedia is less reliable than other reference resources. Furthermore, the rese
arch found at least five unattributed direct quotations and verbatim text from other sources with no citat
ions. More research must be undertaken to analyze Wikipedia entries in other disciplines in order to judge
 the source\'s accuracy and overall quality. This paper also shows the need for analysis of Wikipedia arti
cles\' histories and editing process. This research provides a methodology for further content analysis of
 Wikipedia articles. Although generalizations cannot be made from this paper alone, the paper provides emp
irical data to support concerns regarding the accuracy and authoritativeness of Wikipedia.","Comprehensive
ness,Reliability","Library science"\n"Computational trust in web content quality: a comparative evaluation
 on the Wikipedia project","Pierpaolo Dondio,Stephen Barrett",2007,Informatica,"The problem of identifying
 useful and trustworthy information on the World Wide Web is becoming increasingly acute as new tools such
 as wikis and blogs simplify and democratize publication. It is not hard to predict that in the future the
 direct reliance on this material will expand and the problem of evaluating the trustworthiness of this ki
nd of content become crucial. The Wikipedia project represents the most successful and discussed example o
f such online resources. In this paper we present a method to predict Wikipedia articles trustworthiness b
ased on computational trust techniques and a deep domain-specific analysis. Our assumption is that a deepe
r understanding of what in general defines high-standard and expertise in domains related to Wikipedia h
 i.e. content quality in a collaborative environment h mapped onto Wikipedia elements would lead to a comp
lete set of mechanisms to sustain trust in Wikipedia context. We present a series of experiment. The first
 is a study-case over a specific category of articles; the second is an evaluation over 8 000 articles rep
resenting 65\\% of the overall Wikipedia editing activity. We report encouraging results on the automated
 evaluation of Wikipedia content using our domain-specific expertise method. Finally, in order to appraise
 the value added by using domain-specific expertise, we compare our results with the ones obtained with a
 pre-processed cluster analysis, where complex expertise is mostly replaced by training and automatic clas
sification of common features.","Featured articles,Computational estimation of trustworthiness","Informati
on systems"\n"Computing semantic relatedness using Wikipedia-based explicit semantic analysis","Evgeniy Ga
```

```
brilovich,Shaul Markovitch",2007,"IJCAI\'07 Proceedings of the 20th international joint conference on Arti
fical intelligence","Computing semantic relatedness of natural language texts requires access to vast amou
nts of common-sense and domain-specific world knowledge. We propose Explicit Semantic Analysis (ESA), a no
vel method that represents the meaning of texts in a high-dimensional space of concepts derived from Wikip
edia. We use machine learning techniques to explicitly represent the meaning of any text as a weighted vec
tor of Wikipedia-based concepts. Assessing the relatedness of texts in this space amounts to comparing the
 corresponding vectors using conventional metrics (e.g., cosine). Compared with the previous state of the
 art, using ESA results in substantial improvements in correlation of computed relatedness scores with hum
an judgments: from r = 0.56 to 0.75 for individual words and from r = 0.60 to 0.72 for texts. Importantly,
 due to the use of natural concepts, the ESA model is easy to explain to human users.","Semantic relatedne
ss","Computer science"\n"Computing trust from revision history","Honglei Zeng,Maher A. Alhossaini,Li Ding,
Richard Fikes,Deborah L. McGuinness",2006,"Proceedings of the 2006 International Conference on Privacy, Se
curity and Trust: Bridge the Gap Between PST Technologies and Business Services","A new model of distribut
ed, collaborative information evolution is emerging. As exemplified in Wikipedia, online collaborative inf
ormation repositories are being generated, updated, and maintained by a large and diverse community of use
rs. Issues concerning trust arise when content is generated and updated by diverse populations. Since thes
e information repositories are constantly under revision, trust determination is not simply a static proce
ss. In this paper, we explore ways of utilizing the revision history of an article to assess the trustwort
hiness of the article. We then present an experiment where we used this revision history-based trust model
 to assess the trustworthiness of a chain of successive versions of articles in Wikipedia and evaluated th
e assessments produced by the model.","Featured articles,Vandalism,Reader perceptions of credibility,Compu
tational estimation of trustworthiness","Computer science"\n"Confessions of a librarian or: how I learned
 to stop worrying and love Google","Claire B. Gunnels,Amy Sisson",2009,"Community & Junior College Librari
es","Have you ever stopped to think about life before Google? We will make the argument that Google is the
 first manifestation of Web 2.0, of the power and promise of social networking and the ubiquitous wiki. We
 will discuss the positive influence of Google and how Google and other social networking tools afford lib
rarians leading-edge technologies and new opportunities to teach information literacy. Finally, we will in
clude a top seven list of googlesque tools that no librarian should be without.","Knowledge source for sch
olars and librarians,Student information literacy","Library science"\n"Conflict and consensus in the Chine
se version of Wikipedia","Han-Teng Liao",2009,"IEEE Technology and Society Magazine","It is not easy to in
itiate a new language version of Wikipedia. Although anyone can propose a new language version without fin
ancial cost, certain Wikipedia policies for establishing a new language version must be followed [30]. Onc
e approved and created, the new language version needs tools to facilitate writing and reading in the new
 language. Even if a team tackles these technical and linguistic issues, a nascent community has to then d
evelop its own editorial and administrative policies and guidelines, sometimes by translating and ratifyin
g the policies in another language version (usually English). Given that Wikipedia does not impose an univ
ersal set of editorial and administrative policies and guidelines, the cultural and political nature of su
ch communities remains open-ended.","Cultural and linguistic effects on participation",Linguistics\n"Consi
stency without concurrency control in large, dynamic systems","Mihai Letia,Nuno Pregui\xc3\xa7a,Marc Shapi
ro",2010,"ACM SIGOPS Operating Systems Review","Replicas of a commutative replicated data type {(CRDT)} ev
entually converge without any complex concurrency control. We validate the design of a non-trivial {CRDT},
 a replicated sequence, with performance measurements in the context of Wikipedia. Furthermore, we discuss
 how to eliminate a remaining scalability bottleneck: Whereas garbage collection previously required a sys
```

```
tem-wide consensus, here we propose a flexible two-tier architecture and a protocol for migrating between
 tiers. We also discuss how the {CRDT} concept can be generalised, and its limitations.","Other corpus top
ics","Computer science"\n'
```

```
In [6]: first_2000_chars = c[:2000] # Question 3: Extracting first 2000 characters from the string of text
```

```
In [36]: first_2000_chars # Question 3: Displaying first 2000 characters from the string of text
```

```
Out[36]: b',Author(s),Year,"Published in",Abstract,Topic(s),Domain(s)\n"\'Wikipedia, the free encyclopedia\' as a role
         model? Lessons for open innovation from an exploratory examination of the supposedly democratic-anarchic natu
         re of Wikipedia","Gordon M\xc3\xbcller-Seitz,Guido Reger",2010,"International Journal of Technology Managemen
         t","Accounts of open source software (OSS) development projects frequently stress their democratic, sometimes
         even anarchic nature, in contrast to for-profit organisations. Given this observation, our research evaluates
         qualitative data from Wikipedia, a free online encyclopaedia whose development mechanism allegedly resembles
         that of OSS projects. Our research offers contributions to the field of open innovation research with three m
         ajor findings. First, we shed light on Wikipedia as a phenomenon that has received scant attention from manag
         ement scholars to date. Second, we show that OSS-related motivational mechanisms partially apply to Wikipedia
         participants. Third, our exploration of Wikipedia also reveals that its organisational mechanisms are often p
         erceived as bureaucratic by contributors. This finding was unexpected since this type of problem is often ass
         ociated with for-profit organisations. Such a situation risks attenuating the motivation of contributors and
         sheds a critical light on the nature of Wikipedia as a role model for open innovation processes.","Contributo
         r motivation,Policies and governance,Social order","Information systems"\n"A \'resource review\' of Wikipedi
         a","Cormac Lawler",2006,"Counselling & Psychotherapy Research","The article offers information on Wikipedia,
         an online encyclopedia. The articles and definitions published in Wikipedia can be edited. Articles usually s
         tart as a single sentence and they grow over time through collaborative writing and editing. A discussion pag
         e for every article is also provided for people interested in or concerned with the content of that articl
         e.","Miscellaneous topics","Information systems"\n"A '
```

```
In [63]: data_frame = [first_2000_chars] # Question 4: Converting to list data frame
In [73]: list(data_frame) # Question 4: Displaying list data frame
```

```
Out[73]: [b',Author(s),Year,"Published in",Abstract,Topic(s),Domain(s)\n"\'Wikipedia, the free encyclopedia\' as a rol
e model? Lessons for open innovation from an exploratory examination of the supposedly democratic-anarchic na
ture of Wikipedia","Gordon M\xc3\xbcller-Seitz,Guido Reger",2010,"International Journal of Technology Managem
ent","Accounts of open source software (OSS) development projects frequently stress their democratic, sometim
es even anarchic nature, in contrast to for-profit organisations. Given this observation, our research evalua
tes qualitative data from Wikipedia, a free online encyclopaedia whose development mechanism allegedly resemb
les that of OSS projects. Our research offers contributions to the field of open innovation research with thr
ee major findings. First, we shed light on Wikipedia as a phenomenon that has received scant attention from m
anagement scholars to date. Second, we show that OSS-related motivational mechanisms partially apply to Wikip
edia participants. Third, our exploration of Wikipedia also reveals that its organisational mechanisms are of
ten perceived as bureaucratic by contributors. This finding was unexpected since this type of problem is ofte
n associated with for-profit organisations. Such a situation risks attenuating the motivation of contributors
and sheds a critical light on the nature of Wikipedia as a role model for open innovation processes.","Contri
butor motivation,Policies and governance,Social order","Information systems"\n"A \'resource review\' of Wikip
edia","Cormac Lawler",2006,"Counselling & Psychotherapy Research","The article offers information on Wikipedi
a, an online encyclopedia. The articles and definitions published in Wikipedia can be edited. Articles usuall
y start as a single sentence and they grow over time through collaborative writing and editing. A discussion 
page for every article is also provided for people interested in or concerned with the content of that articl
e.","Miscellaneous topics","Information systems"\n"A ']
```

```
In [129]: split = [i.split() for i in data_frame] # Splitting list data frame
```

In [130]:

```
split # Question 4: Displaying splitted list
```

```
Out[130]:  [[b',Author(s),Year,"Published',
             b'in",Abstract,Topic(s),Domain(s)',
             b'"\'Wikipedia,',
             b'the',
             b'free',
             b"encyclopedia'",
             b'as',
             b'a',
             b'role',
             b'model?',
             b'Lessons',
             b'for',
             b'open',
             b'innovation',
             b'from',
             b'an',
             b'exploratory',
             b'examination',
             b'of',
             b'the',
             b'supposedly',
             b'democratic-anarchic',
             b'nature',
             b'of',
             b'Wikipedia","Gordon',
             b'M\xc3\xbcller-Seitz,Guido',
             b'Reger",2010,"International',
             b'Journal',
             b'of',
             b'Technology',
             b'Management","Accounts',
             b'of',
             b'open',
             b'source',
             b'software',
             b'(OSS)',
             b'development',
             b'projects',
             b'frequently',
             b'stress',
             b'their',
             b'democratic,',
             b'sometimes',
             b'even',
             b'anarchic',
             b'nature,',
             b'in',
             b'contrast',
             b'to',
             b'for-profit',
             b'organisations.',
             b'Given',
             b'this',
             b'observation,',
             b'our',
             b'research',
             b'evaluates',
             b'qualitative',
             b'data',
             b'from',
             b'Wikipedia,',
             b'a',
             b'free',
             b'online',
             b'encyclopaedia',
             b'whose',
             b'development',
             b'mechanism',
             b'allegedly',
             b'resembles',
             b'that',
             b'of',
             b'OSS',
             b'projects.',
             b'Our',
             b'research',
             b'offers',
             b'contributions',
             b'to',
             b'the',
             b'field',
             b'of',
             b'open',
             b'innovation',
             b'research',
             b'with',
             b'three',
             b'major',
             b'findings.',
             b'First,',
             b'we',
             b'shed',
             b'light',
             b'on',
             b'Wikipedia',
             b'as',
             b'a',
             b'phenomenon',
             b'that',
             b'has',
             b'received',
             b'scant',
             b'attention',
             b'from',
             b'management',
             b'scholars',
             b'to',
             b'date.',
             b'Second,',
             b'we',
             b'show',
             b'that',
             b'OSS-related',
             b'motivational',
             b'mechanisms',
             b'partially',
             b'apply',
             b'to',
             b'Wikipedia',
             b'participants.',
             b'Third,',
             b'our',
             b'exploration',
             b'of',
             b'Wikipedia',
             b'also',
             b'reveals',
             b'that',
             b'its',
             b'organisational',
             b'mechanisms',
             b'are',
             b'often',
             b'perceived',
             b'as',
             b'bureaucratic',
             b'by',
             b'contributors.',
             b'This',
             b'finding',
             b'was',
             b'unexpected',
             b'since',
             b'this',
             b'type',
             b'of',
             b'problem',
             b'is',
             b'often',
             b'associated',
             b'with',
             b'for-profit',
             b'organisations.',
             b'Such',
             b'a',
             b'situation',
             b'risks',
             b'attenuating',
             b'the',
             b'motivation',
             b'of',
             b'contributors',
             b'and',
             b'sheds',
             b'a',
             b'critical',
             b'light',
             b'on',
             b'the',
             b'nature',
             b'of',
             b'Wikipedia',
             b'as',
             b'a',
             b'role',
             b'model',
             b'for',
             b'open',
             b'innovation',
             b'processes.","Contributor',
             b'motivation,Policies',
             b'and',
             b'governance,Social',
             b'order","Information',
             b'systems"',
             b'"A',
             b"'resource",
             b"review'",
             b'of',
             b'Wikipedia","Cormac',
             b'Lawler",2006,"Counselling',
             b'&',
             b'Psychotherapy',
             b'Research","The',
             b'article',
             b'offers',
             b'information',
             b'on',
             b'Wikipedia,',
             b'an',
             b'online',
             b'encyclopedia.',
             b'The',
             b'articles',
             b'and',
             b'definitions',
             b'published',
             b'in',
             b'Wikipedia',
             b'can',
             b'be',
             b'edited.',
             b'Articles',
             b'usually',
             b'start',
             b'as',
             b'a',
             b'single',
             b'sentence',
             b'and',
             b'they',
             b'grow',
             b'over',
             b'time',
             b'through',
             b'collaborative',
             b'writing',
             b'and',
             b'editing.',
             b'A',
             b'discussion',
             b'page',
             b'for',
             b'every',
             b'article',
             b'is',
             b'also',
             b'provided',
             b'for',
             b'people',
             b'interested',
             b'in',
             b'or',
             b'concerned',
             b'with',
             b'the',
             b'content',
             b'of',
             b'that',
             b'article.","Miscellaneous',
             b'topics","Information',
             b'systems"',
             b'"A']]
```

In [131]:

```
# Question 4: Counting the occurences of all words in the data frame

from collections import Counter

split = [b',Author(s),Year,"Published',
  b'in",Abstract,Topic(s),Domain(s)',
  b'"\'Wikipedia,',
  b'the',
  b'free',
  b"encyclopedia'",
  b'as',
  b'a',
  b'role',
  b'model?',
  b'Lessons',
  b'for',
  b'open',
  b'innovation',
  b'from',
  b'an',
  b'exploratory',
  b'examination',
  b'of',
  b'the',
  b'supposedly',
  b'democratic-anarchic',
  b'nature',
  b'of',
  b'Wikipedia","Gordon',
  b'M\xc3\xbcller-Seitz,Guido',
  b'Reger",2010,"International',
  b'Journal',
  b'of',
  b'Technology',
  b'Management","Accounts',
  b'of',
  b'open',
  b'source',
  b'software',
  b'(OSS)',
  b'development',
  b'projects',
  b'frequently',
  b'stress',
  b'their',
  b'democratic,',
  b'sometimes',
  b'even',
  b'anarchic',
  b'nature,',
  b'in',
  b'contrast',
  b'to',
  b'for-profit',
  b'organisations.',
  b'Given',
  b'this',
  b'observation,',
  b'our',
  b'research',
  b'evaluates',
  b'qualitative',
  b'data',
  b'from',
  b'Wikipedia,',
  b'a',
  b'free',
  b'online',
  b'encyclopaedia',
  b'whose',
  b'development',
  b'mechanism',
  b'allegedly',
  b'resembles',
  b'that',
  b'of',
  b'OSS',
  b'projects.',
  b'Our',
  b'research',
  b'offers',
  b'contributions',
  b'to',
  b'the',
  b'field',
  b'of',
  b'open',
  b'innovation',
  b'research',
  b'with',
  b'three',
  b'major',
  b'findings.',
  b'First,',
  b'we',
  b'shed',
  b'light',
  b'on',
  b'Wikipedia',
  b'as',
  b'a',
  b'phenomenon',
  b'that',
  b'has',
  b'received',
  b'scant',
  b'attention',
  b'from',
  b'management',
  b'scholars',
  b'to',
  b'date.',
  b'Second,',
  b'we',
  b'show',
  b'that',
  b'OSS-related',
  b'motivational',
  b'mechanisms',
  b'partially',
  b'apply',
  b'to',
  b'Wikipedia',
  b'participants.',
  b'Third,',
  b'our',
  b'exploration',
  b'of',
  b'Wikipedia',
  b'also',
  b'reveals',
  b'that',
  b'its',
  b'organisational',
  b'mechanisms',
  b'are',
  b'often',
  b'perceived',
  b'as',
  b'bureaucratic',
  b'by',
  b'contributors.',
  b'This',
  b'finding',
  b'was',
  b'unexpected',
  b'since',
  b'this',
  b'type',
  b'of',
  b'problem',
  b'is',
  b'often',
  b'associated',
  b'with',
  b'for-profit',
  b'organisations.',
  b'Such',
  b'a',
  b'situation',
  b'risks',
  b'attenuating',
  b'the',
  b'motivation',
  b'of',
  b'contributors',
  b'and',
  b'sheds',
  b'a',
  b'critical',
  b'light',
  b'on',
  b'the',
  b'nature',
  b'of',
  b'Wikipedia',
  b'as',
  b'a',
  b'role',
  b'model',
  b'for',
  b'open',
  b'innovation',
  b'processes.","Contributor',
  b'motivation,Policies',
  b'and',
  b'governance,Social',
  b'order","Information',
  b'systems"',
  b'"A',
  b"'resource",
  b"review'",
  b'of',
  b'Wikipedia","Cormac',
  b'Lawler",2006,"Counselling',
  b'&',
  b'Psychotherapy',
  b'Research","The',
  b'article',
  b'offers',
  b'information',
  b'on',
  b'Wikipedia,',
  b'an',
  b'online',
  b'encyclopedia.',
  b'The',
  b'articles',
  b'and',
  b'definitions',
  b'published',
  b'in',
  b'Wikipedia',
  b'can',
  b'be',
  b'edited.',
  b'Articles',
  b'usually',
  b'start',
  b'as',
  b'a',
  b'single',
  b'sentence',
  b'and',
  b'they',
  b'grow',
  b'over',
  b'time',
  b'through',
  b'collaborative',
  b'writing',
  b'and',
  b'editing.',
  b'A',
  b'discussion',
  b'page',
  b'for',
  b'every',
  b'article',
  b'is',
  b'also',
  b'provided',
  b'for',
  b'people',
  b'interested',
  b'in',
  b'or',
  b'concerned',
  b'with',
  b'the',
  b'content',
  b'of',
  b'that',
  b'article.","Miscellaneous',
  b'topics","Information',
  b'systems"',
  b'"A']
Counter(split)
```

```
Out[131]: Counter({b'"\'Wikipedia,': 1,
                   b'"A': 2,
                   b'&': 1,
                   b"'resource": 1,
                   b'(OSS)': 1,
                   b',Author(s),Year,"Published': 1,
                   b'A': 1,
                   b'Articles': 1,
                   b'First,': 1,
                   b'Given': 1,
                   b'Journal': 1,
                   b'Lawler",2006,"Counselling': 1,
                   b'Lessons': 1,
                   b'Management","Accounts': 1,
                   b'M\xc3\xbcller-Seitz,Guido': 1,
                   b'OSS': 1,
                   b'OSS-related': 1,
                   b'Our': 1,
                   b'Psychotherapy': 1,
                   b'Reger",2010,"International': 1,
                   b'Research","The': 1,
                   b'Second,': 1,
                   b'Such': 1,
                   b'Technology': 1,
                   b'The': 1,
                   b'Third,': 1,
                   b'This': 1,
                   b'Wikipedia': 5,
                   b'Wikipedia","Cormac': 1,
                   b'Wikipedia","Gordon': 1,
                   b'Wikipedia,': 2,
                   b'a': 7,
                   b'allegedly': 1,
                   b'also': 2,
                   b'an': 2,
                   b'anarchic': 1,
                   b'and': 5,
                   b'apply': 1,
                   b'are': 1,
                   b'article': 2,
                   b'article.","Miscellaneous': 1,
                   b'articles': 1,
                   b'as': 5,
                   b'associated': 1,
                   b'attention': 1,
                   b'attenuating': 1,
                   b'be': 1,
                   b'bureaucratic': 1,
                   b'by': 1,
                   b'can': 1,
                   b'collaborative': 1,
                   b'concerned': 1,
                   b'content': 1,
                   b'contrast': 1,
                   b'contributions': 1,
                   b'contributors': 1,
                   b'contributors.': 1,
                   b'critical': 1,
                   b'data': 1,
                   b'date.': 1,
                   b'definitions': 1,
                   b'democratic,': 1,
                   b'democratic-anarchic': 1,
                   b'development': 2,
                   b'discussion': 1,
                   b'edited.': 1,
                   b'editing.': 1,
                   b'encyclopaedia': 1,
                   b"encyclopedia'": 1,
                   b'encyclopedia.': 1,
                   b'evaluates': 1,
                   b'even': 1,
                   b'every': 1,
                   b'examination': 1,
                   b'exploration': 1,
                   b'exploratory': 1,
                   b'field': 1,
                   b'finding': 1,
                   b'findings.': 1,
                   b'for': 4,
                   b'for-profit': 2,
                   b'free': 2,
                   b'frequently': 1,
                   b'from': 3,
                   b'governance,Social': 1,
                   b'grow': 1,
                   b'has': 1,
                   b'in': 3,
                   b'in",Abstract,Topic(s),Domain(s)': 1,
                   b'information': 1,
                   b'innovation': 3,
                   b'interested': 1,
                   b'is': 2,
                   b'its': 1,
                   b'light': 2,
                   b'major': 1,
                   b'management': 1,
                   b'mechanism': 1,
                   b'mechanisms': 2,
                   b'model': 1,
                   b'model?': 1,
                   b'motivation': 1,
                   b'motivation,Policies': 1,
                   b'motivational': 1,
                   b'nature': 2,
                   b'nature,': 1,
                   b'observation,': 1,
                   b'of': 12,
                   b'offers': 2,
                   b'often': 2,
                   b'on': 3,
                   b'online': 2,
                   b'open': 4,
                   b'or': 1,
                   b'order","Information': 1,
                   b'organisational': 1,
                   b'organisations.': 2,
                   b'our': 2,
                   b'over': 1,
                   b'page': 1,
                   b'partially': 1,
                   b'participants.': 1,
                   b'people': 1,
                   b'perceived': 1,
                   b'phenomenon': 1,
                   b'problem': 1,
                   b'processes.","Contributor': 1,
                   b'projects': 1,
                   b'projects.': 1,
                   b'provided': 1,
                   b'published': 1,
                   b'qualitative': 1,
                   b'received': 1,
                   b'research': 3,
                   b'resembles': 1,
                   b'reveals': 1,
                   b"review'": 1,
                   b'risks': 1,
                   b'role': 2,
                   b'scant': 1,
                   b'scholars': 1,
                   b'sentence': 1,
                   b'shed': 1,
                   b'sheds': 1,
                   b'show': 1,
                   b'since': 1,
                   b'single': 1,
                   b'situation': 1,
                   b'software': 1,
                   b'sometimes': 1,
                   b'source': 1,
                   b'start': 1,
                   b'stress': 1,
                   b'supposedly': 1,
                   b'systems"': 2,
                   b'that': 5,
                   b'the': 6,
                   b'their': 1,
                   b'they': 1,
                   b'this': 2,
                   b'three': 1,
                   b'through': 1,
                   b'time': 1,
                   b'to': 4,
                   b'topics","Information': 1,
                   b'type': 1,
                   b'unexpected': 1,
                   b'usually': 1,
                   b'was': 1,
                   b'we': 2,
                   b'whose': 1,
                   b'with': 3,
                   b'writing': 1})
```

In [148]:

```
# Question 5: Counting the "most common" occurences of words in the data frame

from collections import Counter

split = [b',Author(s),Year,"Published',
  b'in",Abstract,Topic(s),Domain(s)',
  b'"\'Wikipedia,',
  b'the',
  b'free',
  b"encyclopedia'",
  b'as',
  b'a',
  b'role',
  b'model?',
  b'Lessons',
  b'for',
  b'open',
  b'innovation',
  b'from',
  b'an',
  b'exploratory',
  b'examination',
  b'of',
  b'the',
  b'supposedly',
  b'democratic-anarchic',
  b'nature',
  b'of',
  b'Wikipedia","Gordon',
  b'M\xc3\xbcller-Seitz,Guido',
  b'Reger",2010,"International',
  b'Journal',
  b'of',
  b'Technology',
  b'Management","Accounts',
  b'of',
  b'open',
  b'source',
  b'software',
  b'(OSS)',
  b'development',
  b'projects',
  b'frequently',
  b'stress',
  b'their',
  b'democratic,',
  b'sometimes',
  b'even',
  b'anarchic',
  b'nature,',
  b'in',
  b'contrast',
  b'to',
  b'for-profit',
  b'organisations.',
  b'Given',
  b'this',
  b'observation,',
  b'our',
  b'research',
  b'evaluates',
  b'qualitative',
  b'data',
  b'from',
  b'Wikipedia,',
  b'a',
  b'free',
  b'online',
  b'encyclopaedia',
  b'whose',
  b'development',
  b'mechanism',
  b'allegedly',
  b'resembles',
  b'that',
  b'of',
  b'OSS',
  b'projects.',
  b'Our',
  b'research',
  b'offers',
  b'contributions',
  b'to',
  b'the',
  b'field',
  b'of',
  b'open',
  b'innovation',
  b'research',
  b'with',
  b'three',
  b'major',
  b'findings.',
  b'First,',
  b'we',
  b'shed',
  b'light',
  b'on',
  b'Wikipedia',
  b'as',
  b'a',
  b'phenomenon',
  b'that',
  b'has',
  b'received',
  b'scant',
  b'attention',
  b'from',
  b'management',
  b'scholars',
  b'to',
  b'date.',
  b'Second,',
  b'we',
  b'show',
  b'that',
  b'OSS-related',
  b'motivational',
  b'mechanisms',
  b'partially',
  b'apply',
  b'to',
  b'Wikipedia',
  b'participants.',
  b'Third,',
  b'our',
  b'exploration',
  b'of',
  b'Wikipedia',
  b'also',
  b'reveals',
  b'that',
  b'its',
  b'organisational',
  b'mechanisms',
  b'are',
  b'often',
  b'perceived',
  b'as',
  b'bureaucratic',
  b'by',
  b'contributors.',
  b'This',
  b'finding',
  b'was',
  b'unexpected',
  b'since',
  b'this',
  b'type',
  b'of',
  b'problem',
  b'is',
  b'often',
  b'associated',
  b'with',
  b'for-profit',
  b'organisations.',
  b'Such',
  b'a',
  b'situation',
  b'risks',
  b'attenuating',
  b'the',
  b'motivation',
  b'of',
  b'contributors',
  b'and',
  b'sheds',
  b'a',
  b'critical',
  b'light',
  b'on',
  b'the',
  b'nature',
  b'of',
  b'Wikipedia',
  b'as',
  b'a',
  b'role',
  b'model',
  b'for',
  b'open',
  b'innovation',
  b'processes.","Contributor',
  b'motivation,Policies',
  b'and',
  b'governance,Social',
  b'order","Information',
  b'systems"',
  b'"A',
  b"'resource",
  b"review'",
  b'of',
  b'Wikipedia","Cormac',
  b'Lawler",2006,"Counselling',
  b'&',
  b'Psychotherapy',
  b'Research","The',
  b'article',
  b'offers',
  b'information',
  b'on',
  b'Wikipedia,',
  b'an',
  b'online',
  b'encyclopedia.',
  b'The',
  b'articles',
  b'and',
  b'definitions',
  b'published',
  b'in',
  b'Wikipedia',
  b'can',
  b'be',
  b'edited.',
  b'Articles',
  b'usually',
  b'start',
  b'as',
  b'a',
  b'single',
  b'sentence',
  b'and',
  b'they',
  b'grow',
  b'over',
  b'time',
  b'through',
  b'collaborative',
  b'writing',
  b'and',
  b'editing.',
  b'A',
  b'discussion',
  b'page',
  b'for',
  b'every',
  b'article',
  b'is',
  b'also',
  b'provided',
  b'for',
  b'people',
  b'interested',
  b'in',
  b'or',
  b'concerned',
  b'with',
  b'the',
  b'content',
  b'of',
  b'that',
  b'article.","Miscellaneous',
  b'topics","Information',
  b'systems"',
  b'"A']
```

```
Counter(split).most_common
```

Out[148]:

```
<bound method Counter.most_common of Counter({b'of': 12, b'a': 7, b'the': 6, b'as': 5, b'that': 5, b'Wikipedi
a': 5, b'and': 5, b'for': 4, b'open': 4, b'to': 4, b'innovation': 3, b'from': 3, b'in': 3, b'research': 3,
b'with': 3, b'on': 3, b'free': 2, b'role': 2, b'an': 2, b'nature': 2, b'development': 2, b'for-profit': 2,
b'organisations.': 2, b'this': 2, b'our': 2, b'Wikipedia,': 2, b'online': 2, b'offers': 2, b'we': 2, b'ligh
t': 2, b'mechanisms': 2, b'also': 2, b'often': 2, b'is': 2, b'systems"': 2, b'"A': 2, b'article': 2, b',Autho
r(s),Year,"Published': 1, b'in",Abstract,Topic(s),Domain(s)': 1, b'"\'Wikipedia,': 1, b"encyclopedia'": 1,
b'model?': 1, b'Lessons': 1, b'exploratory': 1, b'examination': 1, b'supposedly': 1, b'democratic-anarchic':
1, b'Wikipedia","Gordon': 1, b'M\xc3\xbcller-Seitz,Guido': 1, b'Reger",2010,"International': 1, b'Journal':
1, b'Technology': 1, b'Management","Accounts': 1, b'source': 1, b'software': 1, b'(OSS)': 1, b'projects': 1,
b'frequently': 1, b'stress': 1, b'their': 1, b'democratic,': 1, b'sometimes': 1, b'even': 1, b'anarchic': 1,
b'nature,': 1, b'contrast': 1, b'Given': 1, b'observation,': 1, b'evaluates': 1, b'qualitative': 1, b'data':
1, b'encyclopaedia': 1, b'whose': 1, b'mechanism': 1, b'allegedly': 1, b'resembles': 1, b'OSS': 1, b'project
s.': 1, b'Our': 1, b'contributions': 1, b'field': 1, b'three': 1, b'major': 1, b'findings.': 1, b'First,': 1,
b'shed': 1, b'phenomenon': 1, b'has': 1, b'received': 1, b'scant': 1, b'attention': 1, b'management': 1, b'sc
holars': 1, b'date.': 1, b'Second,': 1, b'show': 1, b'OSS-related': 1, b'motivational': 1, b'partially': 1,
b'apply': 1, b'participants.': 1, b'Third,': 1, b'exploration': 1, b'reveals': 1, b'its': 1, b'organisationa
l': 1, b'are': 1, b'perceived': 1, b'bureaucratic': 1, b'by': 1, b'contributors.': 1, b'This': 1, b'finding':
1, b'was': 1, b'unexpected': 1, b'since': 1, b'type': 1, b'problem': 1, b'associated': 1, b'Such': 1, b'situa
tion': 1, b'risks': 1, b'attenuating': 1, b'motivation': 1, b'contributors': 1, b'sheds': 1, b'critical': 1,
b'model': 1, b'processes.","Contributor': 1, b'motivation,Policies': 1, b'governance,Social': 1, b'order","In
formation': 1, b"'resource": 1, b"review'": 1, b'Wikipedia","Cormac': 1, b'Lawler",2006,"Counselling': 1,
b'&': 1, b'Psychotherapy': 1, b'Research","The': 1, b'information': 1, b'encyclopedia.': 1, b'The': 1, b'arti
cles': 1, b'definitions': 1, b'published': 1, b'can': 1, b'be': 1, b'edited.': 1, b'Articles': 1, b'usually':
1, b'start': 1, b'single': 1, b'sentence': 1, b'they': 1, b'grow': 1, b'over': 1, b'time': 1, b'through': 1,
b'collaborative': 1, b'writing': 1, b'editing.': 1, b'A': 1, b'discussion': 1, b'page': 1, b'every': 1, b'pro
vided': 1, b'people': 1, b'interested': 1, b'or': 1, b'concerned': 1, b'content': 1, b'article.","Miscellaneo
us': 1, b'topics","Information': 1})>
```

In [146]:

```
# Question 6: Costruct a bag of matrix -- This will lowercase everthing, and ignore all punctuation by defaul
t
# It will also remove stop words

from sklearn.feature_extraction.text import CountVectorizer

vectorizer = CountVectorizer(lowercase=True, stop_words="english")
```

In [151]:

```
matrix = vectorizer.fit_transform(split) # Question 6: Bag of words matrix
```

In [152]:

```
print(matrix.todense()) # Question 6: Bag of words matrix
```

```
[[0 0 0 ... 0 0 1]
 [0 0 1 ... 0 0 0]
 [0 0 0 ... 1 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```