COMP8210 — Big Data Technologies

Week 8 Lecture 1: Analysing Unstructured Data

Diego Mollá

COMP8210 2021H2

Abstract

In this lecture we will introduce some of the key concepts about analysing unstructured data, with special emphasis on text data.

Update September 21, 2021

Contents

Reading

- Lecture notes.

1 Analysing Unstructured Data

Why Analyse Unstructured Data

It's about Variety

- Probably the biggest impact of Big Data in companies is the possibility to analyse unstructured data.
- Unstructured data contains information that can potentially be very useful.
- It opens up the possibility to access yet untapped information from multiple sources.

Sources of Unstructured Data

Video: surveillance cameras, videos in social media.

Images: Web images, images in social media, satellite images.

Sound: Call centre recordings.

Text: Documents, reports, webpages, social media posts.

Motivating Example: United Healthcare
Davenport & Dyche (2013). Big Data in Big Companies

- Have recorded voice files from customer calls to call centres.
- The voice data was converted to text using speech-to-text conversion tools.
- The text was then analysed using natural language processing software.
- Their analysis focused on identifying customers who use terms suggesting strong dissatisfaction.
- A United representative can then make some sort of intervention.

From T. Davenport & J. Dyché, "Big Data in Big Companies", International Institute for Analytics, 2013, p8.

United Healthcare, like many large organizations pursuing big data, has been focused on structured data analysis for many years, and even advertises its analytical capabilities to consumers ("Health in Numbers"). Now, however, it is focusing its analytical attention on unstructured data — in particular, the data on customer attitudes that is sitting in recorded voice files from customer calls to call centers. The level of customer satisfaction is increasingly important to health insurers, because consumers increasingly have choice about what health plans they belong to. Service levels are also being monitored by state and federal government groups, and published by organizations such as Consumer Reports.

In the past, that valuable data from calls couldn't be analyzed. Now, however, United is turning the voice data into text, and then analyzing it with "natural language processing" software. The analysis process can identify — though it's not easy, given the vagaries of the English language — customers who use terms suggesting strong dissatisfaction. A United representative can then make some sort of intervention — perhaps a call exploring the nature of the problem. The decision being made is the same as in the past —how to identify a dissatisfied customer —but the tools are different.

To analyze the text data, United Healthcare uses a variety of tools. The data initially goes into a "data lake" using Hadoop and NoSQL storage, so the data doesn't have to be normalized. The natural language processing — primarily a "singular value decomposition", or modified word count — takes place on a database appliance. A variety of other technologies are being surveyed and tested to assess their fit within the "future state architecture". United also makes use of interfaces between its statistical analysis tools and Hadoop.

The work to put the customer satisfaction data, along with many other sources of customer data, into a customer data warehouse and analyze it is being led by Mark Pitts, who is based in the Finance organization. However, several other functions and units of United, including its Optum business specializing in selling data and related services to healthcare organizations, are participating. Pitt's team includes both conventional quantitative analysts and data scientists with strong IT and data management skills.

Use Cases of Image Analytics

1. Identify bags at airports.
2. Analyse social media images for missing persons.
3. Real-time vehicle damage assessment.
4. Detect pneumonia from chest x-rays.

Extract from blog post *https://www.zencos.com/blog/5-amazing-use-cases-of-image-analytics/*:
Posted on 13/6/2018
The applications of image analytics are endless. Companies are starting to realize the possibilities of how to extract value from unstructured data. Using images or videos, they can create a new and enticing customer experience within the retail, entertainment, insurance claims, and more.

Here are five image analytics applications that are unexpected, disruptive, and creative.

1. Recognize that Face? Identifying Celebrity Guests at the Royal Wedding

   Curious to know who attended the Royal Wedding? Sky News partnered with Amazon.com and engineering firms to identify the attendees of Prince Harry and Meghan Markle's wedding. They identified celebrity guests using real-time AI capabilities of Amazon Recognition to compare and process live video footage of the guests entering the chapel with known, archived images of celebrity faces.

   In addition to identifying celebrities at a Royal Wedding, image analytics is being used to support an ever-growing list of business use cases. These practical applications of deep learning and image analytics are possible because of the advances in machine learning algorithms, the availability of Big Data, and the existence of robust technology and infrastructure to support real-time processing of these models. Because of these advances, image analytics is now a realistic possibility for a growing number of organizations.

2. Image Analytics Speeds Up Airport Traffic

   According to USA Today, TSA is investing in new scanners that allow agents to "virtually unpack bags." These scanners would provide more accurate object detection, reduce the number of bags that would need to be opened and inspected, and provide faster security screenings. Many U.S. airports are acquiring upgraded technology that enables the use of biometrics such as finger or iris scanning, as an alternative security screening measure. Singapore's Changi Airport will soon be opening a new terminal with automated face recognition. Since the Changi Airport is considered to be "the 6th busiest airport for international traffic," this use of image analytics technology is expected to improve the airport's ability to move passengers from arrival through security to their departure gate and significantly boost capacity.

3. Analyzing Social Media Images for Missing Persons

   Social Media platforms such as Facebook and Google Photos have utilized deep learning and facial recognition for a while now — and, whether you realize it not, all of us who use these platforms are helping them improve the accuracy of their models. If you've ever tagged a friend or family member, you've contributed to refining the model's ability to detect individuals in the photos you post.

   Facial recognition technology is also being used in Australia to identify missing persons. The Missing Persons Action Network (MPAN) is leveraging Facebook as a quick way to spread a message through be-friending missing persons to expand their network through Invisible Friends. With Facebook's facial recognition algorithms, people can be identified in the background of photos. Because of how interconnected friends of friends' networks are, it becomes more possible to find missing persons.

4. Using Image Analytics for Real-time Vehicle Damage Assessment

   Have you ever gotten into a car accident and had to go through a week-long claims process? Wouldn't it be much easier — and less stressful — to be able to pull out your phone, take a few pictures of the damage, and upload them to an app for a real-time assessment? Some insurance companies such as Mitchell Insurance are already using AI technology for automated vehicle damage analysis, allowing for more consistent and timely cost estimates.

5. Image Analytics Provides Doctors with a Second Opinion

   There are countless examples of how to apply deep learning to healthcare. Drug manufacturing companies continuously design and test drugs, treatments, and devices through clinical trials. Clinical research can use deep learning with medical imaging such as CAT scans, X-RAYs, and MRIs to help detect the presence or absence of conditions that are visible in images. For example, the CheXNet algorithm developed by a Stanford Machine Learning group can now detect pneumonia from chest x-rays with accuracy exceeding practicing radiologists. This example illustrates how deep learning and AI are instrumental in improving health care practices, preventing false diagnoses, and providing physicians with a second opinion from a model trained on hundreds of thousands of images.

2 Analysing Text Data

Why Analysing Text?

Information Overload

- A lot of information is available as free text.
- The most natural form to write information is through free text.
- A great deal of digital information is available as free text.
- People can read and understand free text easily.
- But it's very hard for machines!

Examples of Using Text for Big Data

Analysis of social media posts

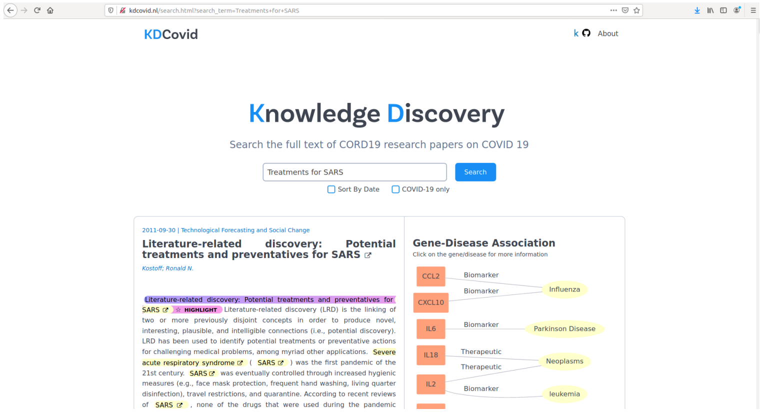
- What do people think about us?
- What do people think about our product?
- What do people think about our competitors?
- What are the most common topics mentioned in media?
- What are the common trends?

Analysis of text documents

- Is this patent claim related to other patents?
- Evidence based medicine: What treatment has best clinical evidence?
- Is this message spam?
- Who is the best person to forward this user request?

Using Text Analytics to Help Combat COVID-19

http://kdcovid.nl/



kdcovid.nl/search.html?search_term=Treatments+for+SARS

KDCovid

k About

Knowledge Discovery

Search the full text of CORD19 research papers on COVID 19

Treatments for SARS Search

□ Sort By Date □ COVID-19 only

2011-09-30 | Technological Forecasting and Social Change

Literature-related discovery: Potential treatments and preventatives for SARS

Kostoff; Ronald N.

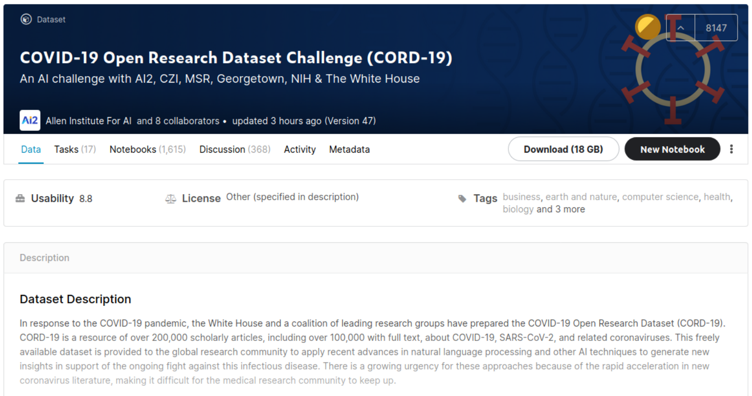
Literature-related discovery: Potential treatments and preventatives for SARS ☆ HIGHLIGHT Literature-related discovery (LRD) is the linking of two or more previously disjoint concepts in order to produce novel, interesting, plausible, and intelligible connections (i.e., potential discovery). LRD has been used to identify potential treatments or preventative actions for challenging medical problems, among myriad other applications. Severe acute respiratory syndrome (SARS) was the first pandemic of the 21st century. SARS was eventually controlled through increased hygienic measures (e.g., face mask protection, frequent hand washing, living quarter disinfection), travel restrictions, and quarantine. According to recent reviews of SARS , none of the drugs that were used during the pandemic

Gene-Disease Association

Click on the gene/disease for more information

The COVID-19 Open Research Dataset Challenge (CORD-19)

https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge



Dataset

COVID-19 Open Research Dataset Challenge (CORD-19)

An AI challenge with AI2, CZI, MSR, Georgetown, NIH & The White House

Ai2 Allen Institute For AI and 8 collaborators • updated 3 hours ago (Version 47)

Usability 8.8

License Other (specified in description)

Tags business, earth and nature, computer science, health, biology and 3 more

Description

Dataset Description

In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 200,000 scholarly articles, including over 100,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in new coronavirus literature, making it difficult for the medical research community to keep up.

2.1 Characteristics of Text

Text as Arbitrary Symbols

- Words are encoded as arbitrary symbols.
- Different languages use different representations to represent the same word.
- Even within one language there is no clear correspondence between a word symbol and its meaning.

https://www.linguisticsociety.org/content/how-many-languages-are-there-world

Ambiguity everywhere

Language features ambiguity at multiple levels.

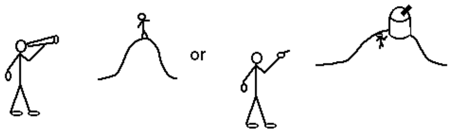
Lexical

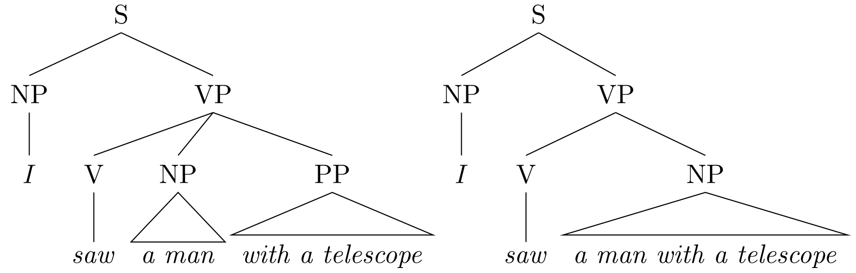
Example from Google's dictionary:

- bank (n): the land alongside or sloping down a river or lake.
- bank (n): financial establishment that uses money deposited by customers for investment, . . .
- bank (v): form in to a mass or mound.
- bank (v): build (a road, railway, or sports track) higher at the outer edge of a bend to facilitate fast cornering.
- . . .

Syntactic

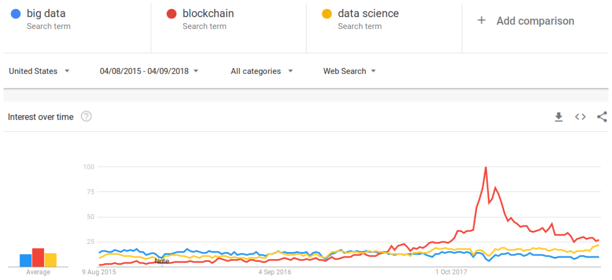
- "I saw a man with a telescope" ... who has the telescope?





So many words!

- Any language features a large number of distinct words.
- New words are coined.
- Words change their use in time.
- There are also names, numbers, dates... the possibilities are infinite.



● big data Search term

● blockchain Search term

● data science Search term

+ Add comparison

United States ▾ 04/08/2015 - 04/09/2018 ▾ All categories ▾ Web Search ▾

Interest over time

https://trends.google.com

2.2 Common Building Blocks for Text Analytics

Tokenisation

- *Tokenisation*: Break down the input into words and other kinds of tokens.
- *Sentence Segmentation*: Break down the input into sentences.
- Tokenisation needs to be done as a first step in other applications.
- Same process as identifying separate units in programming languages, but harder.
- Tokenisation in space-delimited languages is fairly easy but some languages have no clear-cut way to separate words, or even sentences.

Keyword Extraction and Word Clouds

- *Keyword extraction*: Extract the most important words in a document or collection of documents.
- *Word cloud*: a graphical interface that displays words according to their importance.

How to Select and Score words?

- Remove stop words.
- Select words by frequency.
- Use tf.idf
- ...

good
love
everyon
onli
enjoy
like
hope
fun
look
wish
friend
readiguy
veri
work
pretti
day
lol
morn
peopl
long
miss
whi
excit
haha
cool
pleas
movitri
help
final
well
nice
realli
better
best
follow
make
happi
amaz
getlittl
alway
great
awesom
glad
have
thank
watch
right

Removing Stop Words

- Many packages offer lists of stop words.
- These lists include words that usually are not important.
- There is no universal list of stop words.

Stop words in the Python NLTK package

```
>>> from nltk.corpus import stopwords
>>> stopwords.words('english')
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're",
 "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he',
 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's",
 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which',
 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are',
 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do',
 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because',
 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against',
 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below',
 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again',
 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all',
 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no',
 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't',
 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll',
 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn',
 "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven',
 "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't",
 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't",
 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

Selecting Words by Frequency

- If you want to find words that *discriminate* between different documents ...
  - Very frequent words are not useful (because they are in most documents).
  - Very rare words are not useful (because they are in too few documents).
  - The right solution is somewhere in the middle.
- A practical solution is to apply this sequence:
  1. Remove stop words.
  2. Select the most frequent remaining words.

Selecting Words by tf.idf

- tf.idf strikes a balance between words that are frequent but are not too frequent.
- *tf*: Term frequency. Words that are very frequent are more important.

$$tf(w,d) = \text{number of times word } w \text{ occurs in document } d$$

- *idf*: Inverse document frequency. Words that occur in many documents are less important.

$$idf(w) = 1 + \log\left(\frac{\text{number of documents}}{\text{number of documents containing word } w}\right)$$

- $tf.idf(w,d) = tf(w,d) \times idf(w)$
- We select words from document d with high $tf.idf$, possibly after removing stop words.

Stemming and Lemmatisation

- Words in many languages (e.g. English) have inflections.
    - Singular, plural, verb-ing, etc.
- Stemming and lemmatisation allow to group words that are different only because of their inflections.
- *Stemming*: Remove the part of a word that has the inflection to produce the *stem*.
- *Lemmatisation*: Convert an inflected word into a word without inflections to produce the *lemma* or *base form*.
- Stemming is easier and requires less knowledge of the language. Often stemming is all you need.
- Lemmatisation is useful when you want to produce real words.
    - E.g. if you want to display keywords.

Part of Speech Tagging

- Words with the same part of speech have similar grammatical properties.
- In general, one can replace a word with another of the same part of speech and the sentence is still grammatical.
- Most words belong to *open class types*: nouns, verbs, adjectives, adverbs.
    - These words usually determine the topic of the sentence.
    - For example, keywords would normally be words in open class types.
- Words in the *closed class types* are useful to connect other words: prepositions, determiners, pronouns, conjunctions, ...
    - These words are usually removed by some text applications.
    - For example, stop words are normally words from closed class types.

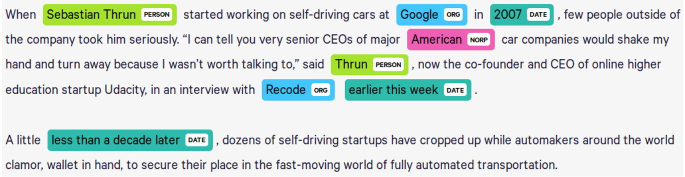
Parts of Speech in the Penn Treebank

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker

MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Named Entity Recognition

- *Named entities* are (often multi-word) expressions that refer to proper names of specific types.
  - Persons, organisations, locations, artifacts, dates, ...
- Named entity recognition is one of the most common tasks in text analytics.



When Sebastian Thrun PERSON started working on self-driving cars at Google ORG in 2007 DATE , few people outside of the company took him seriously. “I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun PERSON , now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE .

A little less than a decade later DATE , dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

https://explosion.ai/demos/displacy-ent

Entities in the Message Understanding Conference

- Named Entities
  - Organization
  - Person
  - Location
- Temporal Expressions
  - Date
  - Time
- Number Expressions
  - Money
  - Percent

MUC

- Initiated and financed by DARPA (Defense Advanced Research Projects Agency).
- From 1987 to 1997.
- The goal was to advance methods for information extraction from text.
- MUC-6 (1995) introduced the task of named entity recognition.
- The MUC named entities have been used by many systems since then.

Text Classification

Many different tasks can be seen as text classification.

- E-mail filtering, spam detection, sentiment analysis . . .

To classify text it needs to be converted into a vector of features.

Feature Selection

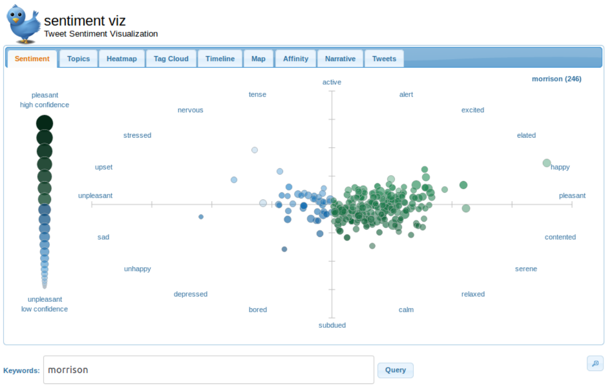
- Extract key words and use them to build document vectors for classification.
- For example, remove *stop words* and/or select words with high tf.idf.

Feature Extraction

- Generate document vectors based on mathematical and statistical combinations of the entire information of the text.
- *Latent Semantic Analysis* (LSA), *Singular Value Decomposition* (SVD) and *Principal Component Analysis* (PCA) are traditionally used for feature extraction.
- More recent approaches use *neural networks* and *word embeddings*.

Sentiment Analysis

- Sentiment analysis is a popular example of text classification.
- Often needed for market analysis.
- A well known approach to analyse social media posts.



Keywords: morrison

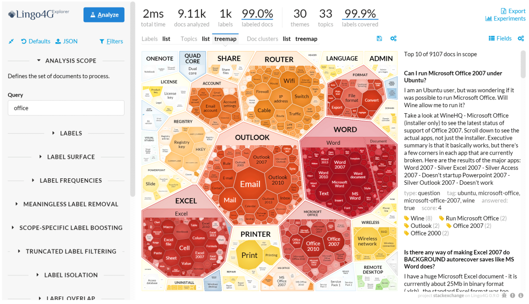
https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Text Retrieval / Filtering

- Often needed to find specific information in large volumes of text.
- Search engines are the first popular applications of text retrieval.
- A common step before doing other processing tasks such as sentiment analysis.

Text Clustering

- Nothing to do with computer clusters . . .
- Useful when we have large volumes of text but no labels.
- Can help characterise types of customers, common views of opinion, etc.



Lingo4G Explorer Analyze

Defaults JSON Filters

ANALYSIS SCOPE

Defines the set of documents to process.

Query: office

LABELS

LABEL SURFACE

LABEL FREQUENCIES

MEANINGLESS LABEL REMOVAL

SCOPE-SPECIFIC LABEL BOOSTING

TRUNCATED LABEL FILTERING

LABEL ISOLATION

LABEL OVERLAP

2ms total time | 9.11k docs analyzed | 1k labels | 99.0% labeled docs | 30 themes | 33 topics | 99.9% labels covered

Labels list | Topics list treemap | Doc clusters list treemap

Export

Experiments

Fields

Top 10 of 9107 docs in scope

Can I run Microsoft Office 2007 under Ubuntu?

I am an Ubuntu user, but was wondering if it was possible to run Microsoft Office. Will Wine allow me to run it?

Take a look at WineHQ - Microsoft Office (installer only) to see the latest status of support of Office 2007. Scroll down to see the actual apps, not just the installer. Executive summary is that it basically works, but there's a few corners in each app that are currently broken. Here are the results of the major apps: Word 2007 - Silver Excel 2007 - Silver Access 2007 - Doesn't startup Powerpoint 2007 - Silver Outlook 2007 - Doesn't work

type: question tag: ubuntu, microsoft-office, microsoft-office-2007, wine answered: true score: 4

Wine (8) Run Microsoft Office (2) Outlook (2) Office 2007 (2) Office 2000 (2)

Is there any way of making Excel 2007 do BACKGROUND autorecover saves like MS Word does?

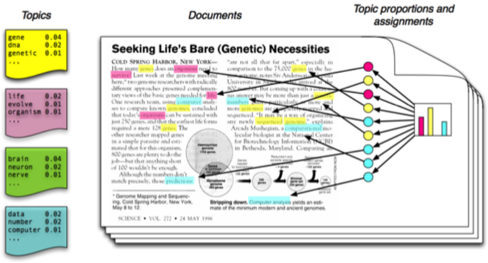
I have a huge Microsoft Excel document - it is currrently about 25Mb in binary format (.xlsb)...the standard Excel format was too

project stackexchange on Lingo4G 0.9.0

https://get.carrotsearch.com/lingo4g/1.4.0/doc/

Topic Modelling

- Topic modelling is a more complex form of unsupervised text processing.
- The task is to find the common topics in a collection of texts (e.g. tweets).
- Implementations such as *Latent Dirichlet Allocation (LDA)* return keywords that are most characteristic of each topic.



Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Take-home Messages

- Sources of unstructured data.
- Impact of unstructured data.
- Characteristics of text.
- Common building blocks for text analytics.

What's Next

Week 9

- Text Analytics (II).