Eksploracja danych internetowych

Ćwiczenie nr 2

Rafał Kowalski, 227148 Arkadiusz Juszczak 227142

2019-05-18

Wstęp

Plan ćwiczenia:

1. Opis wybranej domeny internetowej.
2. Opis procesu przygotowania plików tekstowych zawierających dane o dokumentach.
3. Analiza atrybutów dokumentów w zależności od reprezentacji i ustawienia parametrów w programie "Weka", dla przygotowanych plików tekstowych.
4. Opis otrzymanych wyników analizy klastrowej dokumentów w zalezności od użytych atrybutów.

Cel ćwiczenia:

- Analiza skupień dokumentów ze strony internetowej.

Opis wybranej domeny internetowej

W ćwiczeniu wykorzystano następującą domenę internetową:

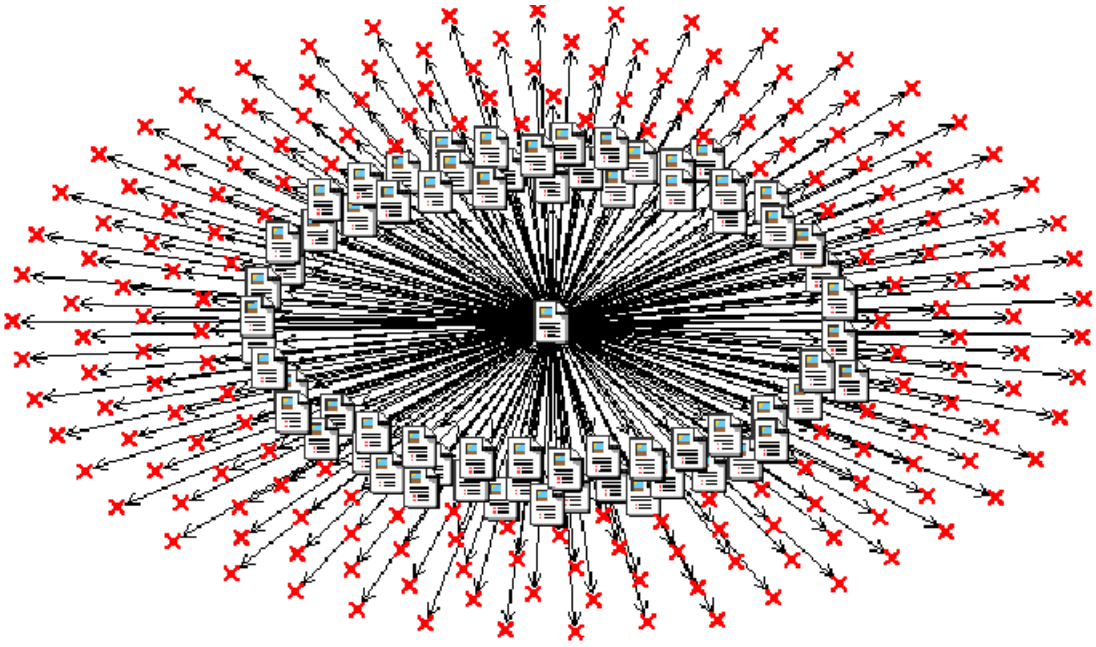
https://arxiv.org/

Domena ta jest prowadzona przez Cornell University. Strona jest bazą danych dla ponad 1.5 miliona przedruków prac naukowych z dziedzin m.in takich jak: fizyki, matematyki, informatyki, statystyki, elektroniki, ekonomii.

Przy pobieraniu stron z domeny zastosowano następującą metodologię

- Wszystkie strony i podstrony zostały pobrane za pomocą programu WebSphinx - crawler.
- Do analizy przyjęto wszystkie linki znajdujące się na stronach.
- Przy pobieraniu uwzględniono pobieranie do głębokości jednej strony (Depth: 1)
- W wyniku konkatenacji uzyskano złączenie w sumie 69 stron internetowych

Domena po przeanalizowaniu posiadała następującą strukturę:



W wynikach analizy widać dużą ilość błędów - błędy te związane są z http 403 - odmowa dostępu lub zastrzeżona zawartość. Pozostałe strony zaznaczone ikonką pliku są dostępne i posłużyły do dalszego procesu.

W wyniku konkatenacji otrzymano jeden plik zawierający w sumie 69 dokumentów - stron internetowych. Nazwano go 1.html

Opis procesu przygotowania plików tekstowych zawierających dane o dokumentach

Przedstawiony plik 1.html musiał zostać poddany wstępnej obróbce. Należało go przekonwertować na plik tekstowy (csv)

Konwersja formatu html do formatu tekstowego:

Przed obróbką przykładowy nagłówek pliku miał postać:

```
<HTML><HEAD><TITLE>Concatenation</TITLE></HEAD><BODY>
<TABLE WIDTH="100%"><TR>
<TD ALIGN=left><A NAME="page1">arXiv.org e-Print archive
[https://arxiv.org/]</A>
<TD ALIGN=right>Page 1</TABLE>
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">



  <link rel="shortcut icon" href="https://arxiv.org/favicon.ico"
type="image/x-icon" />
  <link rel="stylesheet" type="text/css" media="screen"
href="https://static.arxiv.org/css/arXiv.css?v=20190307" />

  <!-- Matomo -->
  <script type="text/javascript">
    var _paq = window._paq || [];
    /* tracker methods like "setCustomDimension" should be called before
"trackPageView" */
    _paq.push(["setCookieDomain", "*.arxiv.org"]);
    _paq.push(['trackPageView']);
    _paq.push(['enableLinkTracking']);
    (function() {
      var u="https://webstats.arxiv.org/";
      _paq.push(['setTrackerUrl', u+'matomo.php']);
      _paq.push(['setSiteId', '1']);
      var d=document, g=d.createElement('script'),
s=d.getElementsByTagName('script')[0];
      g.type='text/javascript'; g.async=true; g.defer=true;
g.src=u+'matomo.js'; s.parentNode.insertBefore(g,s);
    })();
  </script>
  <!-- End Matomo Code -->
  <link rel="stylesheet" href="https://maxcdn.bootstrapcdn.com/font-
awesome/4.7.0/css/font-awesome.min.css">
```

W pliku widoczny jest nagłówek nadany przez program WebSphinx nagłówek o nazwie "CONCATENATION" wskazuje na złączenie kilku dokumentów html w całość. Dodatkowo ustawiono znacznik rozdzielający dokumenty w postaci wyrażenia: "KONIEC":

```
      </div>
    </div>
  </footer>
```

KONIEC

```
<TABLE WIDTH="100%"><TR>
<TD ALIGN=left><A NAME="page2">Log in to arXiv | arXiv e-print repository
[https://arxiv.org/login]</A>
<TD ALIGN=right>Page 2</TABLE>
<!DOCTYPE html>
    <meta charset="utf-8"/>
  <link rel="stylesheet" href="https://maxcdn.bootstrapcdn.com/font-
awesome/4.7.0/css/font-awesome.min.css">
```

Widoczne są więc znaki specjalne oraz wyrażenia i tagi właściwe dla języka Html oraz arkusza stylów CSS. Przerobienie pliku html do postaci tekstowej wymagało przygotowania odpowiedniego skryptu, tak aby wyodrębnić tylko potrzebny tekst w celu dalszej analizy.

Plik 1.html został poddany obróbce za pomocą skryptu napisanego w języku Python 3.7.

```
import html2text
import nltk
import csv

from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from string import digits

# Use following statment at the first run.
nltk.download()

nltk.download('stopwords')
nltk.download('punkt')
stop_words = set(stopwords.words('english'))

documents_content = []

f = open('1.html', 'r')
content = f.read()
divider = 'KONIEC'
documents = content.split(divider)


i = 0
for document in documents:
    print(i)
    i += 1
```

```
    text_maker = html2text.HTML2Text()
    text_maker.ignore_links = True
    text_maker.bypass_tables = False
    text = text_maker.handle(document)
    title = text.split('\n', 1)[0]

    text = text.replace('\n', '')

    special_characters = ['!', '@', '#', '$', '%', '^', '&', '*', '(', ')',
',', '.', ':', ';', '|', '[', ']', '"', '/', 'https://']

    for char in special_characters:
        text = text.replace(char, ' ')
        title = title.replace(char, '')

    remove_digits = str.maketrans('', '', digits)
    text = text.translate(remove_digits)

    text = text.replace('  ', ' ')
    text = text.replace('\t', '')
    title = title.replace('arXiv', '')

    title = title[:50]
    title += str(i)
    documents_content.append([title, text])

with open('res.csv', mode='w') as f:
    writer = csv.writer(
        f,
        delimiter=',',
        quotechar='"',
        quoting=csv.QUOTE_MINIMAL
    )
    id = 1

    keys = ['title', 'content']
    writer.writerow(keys)

    for doc in documents_content:

        writer.writerow(doc)

with open('pages.arff', mode='w') as f:
    s = "@relation dokumenty\n@attribute _title string\n@attribute content
string\n@data\n"
    i = 1
    for title, content in documents_content:
        t = '"{}","{}"\n'.format(title, content)
        i += 1
        s += t

    s = s[:-1]

    f.write(s)
```

Skrypt miał za zadanie:

- usunąć znaczniki (tagi) html
- usunąć znaki specjalne takie jak
- usunąć linki

- usunąć nadmiarowe znaki odstępu (spacje)
- usunąć znaki nowej linii "\n"
- określić koniec i początek nowego dokumentu
- pogrupować plik wynikowy na atrybut opisujący tytuł/numer dokumentu (id) oraz zawartość dokumentu (content)
- zapisać efekt wynikowy do pliku tekstowego typu .csv

W wyniku działania skryptu powstał plik o nazwie res.csv o następującej zawartości (przykładowe 10 pierwszych wierszy):

```
title,content
org e-Print archive httpsarxivorg  Page 11,arXiv org e-Print archive https
arxiv org  Page  https  static arxiv org icons close-slider png  Donate
to arXivPlease join the Simons Foundation and our generous member
organizations insupporting arXiv during our giving campaign  of your
contribution willfund improvements and new initiatives to benefit arXiv's
global scientificcommunity  DONATE  secure site no need to create account
https  webstats arxiv org matomo php?idsite= rec=  Cornell University https
static arxiv org icons cu cornell-reduced-white-SMALL svg We gratefully
acknowledge support from the Simons Foundation and member institutions
arXiv orgLoginAll fields Title Author s Abstract Comments Journal reference
ACMclassification MSC classification Report number arXiv identifier DOI
ORCIDarXiv author ID Help pages Full text Help  Advanced search Open access
to  e-prints in the fields of physics mathematics computer science
quantitative biology quantitative finance statistics electrical engineering
and systems science and economics Submissions toarXiv should conform to
Cornell University academic standards arXiv is ownedand operated by Cornell
University a private not-for-profit educationalinstitution arXiv is funded
by Cornell University the Simons Foundation andby the member institutions
Subject search and browse  Physics Mathematics Quantitative BiologyComputer
Science Quantitative Finance Statistics Electrical Engineering andSystems
Science Economics Apr  Support arXiv with a donation  Jan  The annual
update from the arXiv team is now available  Sept  arXiv looks to the
future with move to Cornell CIS See cumulative What's New pages Read robots
beware before attempting anyautomated download  Physics  Astrophysics
astro-ph  new recent search  includes Astrophysics of Galaxies Cosmology
and Nongalactic Astrophysics Earth and Planetary Astrophysics High Energy
Astrophysical Phenomena Instrumentation and Methods for Astrophysics Solar
and Stellar Astrophysics  Condensed Matter  cond-mat  new recent search
includes Disordered Systems and Neural Networks Materials Science
Mesoscaleand Nanoscale Physics Other Condensed Matter Quantum Gases Soft
CondensedMatter Statistical Mechanics Strongly Correlated Electrons
Superconductivity  General Relativity and Quantum Cosmology  gr-qc  new
recent search  High Energy Physics - Experiment  hep-ex  new recent
search  High Energy Physics - Lattice  hep-lat  new recent search  High
Energy Physics - Phenomenology  hep-ph  new recent search  High Energy
Physics - Theory  hep-th  new recent search  Mathematical Physics  math-
ph  new recent search  Nonlinear Sciences  nlin  new recent search
includes Adaptation and Self-Organizing Systems Cellular Automata
andLattice Gases Chaotic Dynamics Exactly Solvable and Integrable Systems
Pattern Formation and Solitons  Nuclear Experiment  nucl-ex  new recent
search  Nuclear Theory  nucl-th  new recent search  Physics  physics
new recent search  includes Accelerator Physics Applied Physics Atmospheric
and OceanicPhysics Atomic and Molecular Clusters Atomic Physics Biological
Physics Chemical Physics Classical Physics Computational Physics Data
Analysis Statistics and Probability Fluid Dynamics General Physics
Geophysics History and Philosophy of Physics Instrumentation and Detectors
MedicalPhysics Optics Physics and Society Physics Education Plasma Physics
Popular Physics Space Physics  Quantum Physics  quant-ph  new recent
```

```
search  Mathematics  Mathematics  math  new recent search  includes see
detailed description  Algebraic Geometry Algebraic Topology Analysis of
PDEs Category Theory Classical Analysis and ODEs Combinatorics Commutative
Algebra Complex Variables Differential Geometry DynamicalSystems Functional
Analysis General Mathematics General Topology GeometricTopology Group
Theory History and Overview Information Theory K-Theory andHomology Logic
Mathematical Physics Metric Geometry Number Theory Numerical Analysis
Operator Algebras Optimization and Control Probability Quantum Algebra
Representation Theory Rings and Algebras Spectral Theory Statistics Theory
Symplectic Geometry  Computer Science  Computing Research Repository  CoRR
new recent search  includes see detailed description  Artificial
Intelligence Computation andLanguage Computational Complexity Computational
Engineering Finance andScience Computational Geometry Computer Science and
Game Theory ComputerVision and Pattern Recognition Computers and Society
Cryptography andSecurity Data Structures and Algorithms Databases Digital
Libraries Discrete Mathematics Distributed Parallel and Cluster Computing
EmergingTechnologies Formal Languages and Automata Theory General
Literature Graphics Hardware Architecture Human-Computer Interaction
InformationRetrieval Information Theory Logic in Computer Science Machine
Learning Mathematical Software Multiagent Systems Multimedia Networking and
InternetArchitecture Neural and Evolutionary Computing Numerical Analysis
OperatingSystems Other Computer Science Performance Programming Languages
Robotics Social and Information Networks Software Engineering Sound
SymbolicComputation Systems and Control  Quantitative Biology  Quantitative
Biology  q-bio  new recent search  includes see detailed description
Biomolecules Cell Behavior Genomics Molecular Networks Neurons and
Cognition Other Quantitative Biology Populations and Evolution Quantitative
Methods Subcellular Processes Tissues and Organs  Quantitative Finance
Quantitative Finance  q-fin  new recent search  includes see detailed
description  Computational Finance Economics GeneralFinance Mathematical
Finance Portfolio Management Pricing of Securities Risk Management
Statistical Finance Trading and Market Microstructure  Statistics
Statistics  stat  new recent search  includes see detailed description
Applications Computation MachineLearning Methodology Other Statistics
Statistics Theory  Electrical Engineering and Systems Science  Electrical
Engineering and Systems Science  eess  new recent search  includes see
detailed description  Audio and Speech Processing Image andVideo Processing
Signal Processing  Economics  Economics  econ  new recent search  includes
see detailed description  Econometrics General Economics Theoretical
Economics  About arXiv  General information and Scientific Advisory Board
Support and Governance Model and Member Advisory Board  Find view email
alerts and RSS feeds  Submission and moderation details  Usage statistics
and news  See also searchable help pages  About arXiv  Leadership Team
__Contact Us  __Follow us on Twitter  Help  Privacy Policy  Blog
SubscribearXiv® is a registered trademark of Cornell University If you have
a disability and are having trouble accessing information on thiswebsite or
need materials in an alternate format contact web-accessibility cornell edu
for assistance Log in to e-print repository httpsarxivorglogin
```

Jak widać powyżej otrzymany plik nie zawiera już żadnych "nadmiarowych" dodatków. Otrzymany tekst jest "zawartością" danej strony i stanowi podstawę do dalszej analizy.

Konwersja pliku tekstowego .csv do formatu .arff

W celu dokonania analizy za pomocą programu "Weka" wymagano konwersji pliku .csv do formatu .arff

Konwersji dokonano za pomocą skryptu napisanego w języku R.

```
library("foreign")

data=read.csv("res.csv",header=TRUE)
```

```
write.arff(x=data ,file= "pages.arff")
```
W wyniku otrzymano plik pages.arff o następującej strukturze:

```
@relation dokumenty
@attribute _title string
@attribute content string
@data
"org e-Print archive httpsarxivorg  Page 11","arXiv org e-Print archive
https  arxiv org  Page  https  static arxiv org icons close-slider png
Donate to arXivPlease join the Simons Foundation and our generous member
organizations insupporting arXiv during our giving campaign  of your
contribution willfund improvements and new initiatives to benefit arXiv's
global scientificcommunity  DONATE  secure site no need to create account
https  webstats arxiv org matomo php?idsite= rec=  Cornell University https
static arxiv org icons cu cornell-reduced-white-SMALL svg We gratefully
acknowledge support from the Simons Foundation and member institutions
arXiv orgLoginAll fields Title Author s Abstract Comments Journal reference
ACMclassification MSC classification Report number arXiv identifier DOI
ORCIDarXiv author ID Help pages Full text Help  Advanced search Open access
to  e-prints in the fields of physics mathematics computer science
quantitative biology quantitative finance statistics electrical engineering
and systems science and economics Submissions toarXiv should conform to
Cornell University academic standards arXiv is ownedand operated by Cornell
University a private not-for-profit educationalinstitution arXiv is funded
by Cornell University the Simons Foundation andby the member institutions
Subject search and browse  Physics Mathematics Quantitative BiologyComputer
Science Quantitative Finance Statistics Electrical Engineering andSystems
Science Economics Apr  Support arXiv with a donation  Jan  The annual
update from the arXiv team is now available  Sept  arXiv looks to the
future with move to Cornell CIS See cumulative What's New pages Read robots
beware before attempting anyautomated download  Physics  Astrophysics
astro-ph  new recent search  includes Astrophysics of Galaxies Cosmology
and Nongalactic Astrophysics Earth and Planetary Astrophysics High Energy
Astrophysical Phenomena Instrumentation and Methods for Astrophysics Solar
and Stellar Astrophysics  Condensed Matter  cond-mat  new recent search
includes Disordered Systems and Neural Networks Materials Science
Mesoscaleand Nanoscale Physics Other Condensed Matter Quantum Gases Soft
CondensedMatter Statistical Mechanics Strongly Correlated Electrons
Superconductivity  General Relativity and Quantum Cosmology  gr-qc  new
recent search  High Energy Physics - Experiment  hep-ex  new recent
search  High Energy Physics - Lattice  hep-lat  new recent search  High
Energy Physics - Phenomenology  hep-ph  new recent search  High Energy
Physics - Theory  hep-th  new recent search  Mathematical Physics  math-
ph  new recent search  Nonlinear Sciences  nlin  new recent search
includes Adaptation and Self-Organizing Systems Cellular Automata
andLattice Gases Chaotic Dynamics Exactly Solvable and Integrable Systems
Pattern Formation and Solitons  Nuclear Experiment  nucl-ex  new recent
search  Nuclear Theory  nucl-th  new recent search  Physics  physics
new recent search  includes Accelerator Physics Applied Physics Atmospheric
and OceanicPhysics Atomic and Molecular Clusters Atomic Physics Biological
Physics Chemical Physics Classical Physics Computational Physics Data
Analysis Statistics and Probability Fluid Dynamics General Physics
Geophysics History and Philosophy of Physics Instrumentation and Detectors
MedicalPhysics Optics Physics and Society Physics Education Plasma Physics
Popular Physics Space Physics  Quantum Physics  quant-ph  new recent
search  Mathematics  Mathematics  math  new recent search  includes see
detailed description  Algebraic Geometry Algebraic Topology Analysis of
PDEs Category Theory Classical Analysis and ODEs Combinatorics Commutative
Algebra Complex Variables Differential Geometry DynamicalSystems Functional
Analysis General Mathematics General Topology GeometricTopology Group
Theory History and Overview Information Theory K-Theory andHomology Logic
```

Mathematical Physics Metric Geometry Number Theory Numerical Analysis Operator Algebras Optimization and Control Probability Quantum Algebra Representation Theory Rings and Algebras Spectral Theory Statistics Theory Symplectic Geometry Computer Science Computing Research Repository CoRR new recent search includes see detailed description Artificial Intelligence Computation andLanguage Computational Complexity Computational Engineering Finance andScience Computational Geometry Computer Science and Game Theory ComputerVision and Pattern Recognition Computers and Society Cryptography andSecurity Data Structures and Algorithms Databases Digital Libraries Discrete Mathematics Distributed Parallel and Cluster Computing EmergingTechnologies Formal Languages and Automata Theory General Literature Graphics Hardware Architecture Human-Computer Interaction InformationRetrieval Information Theory Logic in Computer Science Machine Learning Mathematical Software Multiagent Systems Multimedia Networking and InternetArchitecture Neural and Evolutionary Computing Numerical Analysis OperatingSystems Other Computer Science Performance Programming Languages Robotics Social and Information Networks Software Engineering Sound SymbolicComputation Systems and Control Quantitative Biology Quantitative Biology q-bio new recent search includes see detailed description Biomolecules Cell Behavior Genomics Molecular Networks Neurons and Cognition Other Quantitative Biology Populations and Evolution Quantitative Methods Subcellular Processes Tissues and Organs Quantitative Finance Quantitative Finance q-fin new recent search includes see detailed description Computational Finance Economics GeneralFinance Mathematical Finance Portfolio Management Pricing of Securities Risk Management Statistical Finance Trading and Market Microstructure Statistics Statistics stat new recent search includes see detailed description Applications Computation MachineLearning Methodology Other Statistics Statistics Theory Electrical Engineering and Systems Science Electrical Engineering and Systems Science eess new recent search includes see detailed description Audio and Speech Processing Image andVideo Processing Signal Processing Economics Economics econ new recent search includes see detailed description Econometrics General Economics Theoretical Economics About arXiv General information and Scientific Advisory Board Support and Governance Model and Member Advisory Board Find view email alerts and RSS feeds Submission and moderation details Usage statistics and news See also searchable help pages About arXiv Leadership Team __Contact Us __Follow us on Twitter Help Privacy Policy Blog SubscribearXiv® is a registered trademark of Cornell University If you have a disability and are having trouble accessing information on thiswebsite or need materials in an alternate format contact web-accessibility cornell edu for assistance "
"Log in to e-print repository httpsarxivorglogin2","Log in to arXiv arXiv e-print repository https arxiv org login Page https webanalytics library cornell edu piwik php?idsite= rec= Cornell University https arxiv org static base images cornell-reduced-white-SMALL svg We gratefully acknowledge support from the Simons Foundation and member institutions arXiv https arxiv org static base images arxiv-logo-web svg Help Advanced SearchAll fields Title Author Abstract Comments Journal reference ACM classificationMSC classification Report number arXiv identifier DOI ORCID arXiv author IDHelp pages Full textSearchLoginAccounts v released -- Log in to arXiv orgIf you're already registeredUsername or e-mailPasswordForgot your password? If you've never logged in to arXiv orgRegister for the first timeRegistration is required to submit or update papers About arXiv Leadership Team __Contact __Follow us on Twitter Help Privacy Policy Blog SubscribearXiv® is a registered trademark of Cornell University If you have a disability and are having trouble accessing information on thiswebsite or need materials in an alternate format contact web-accessibility cornell edu for assistance "
"Search e-print repository httpsarxivorgsearch P3","Search arXiv e-print repository https arxiv org search Page https webstats arxiv org

matomo php?idsite= rec= CornellUniversity https static arxiv org static base images cornell-reduced-white-SMALL svg We gratefully acknowledge support from the Simons Foundation and member institutions arXiv https static arxiv org static base images arxiv-logo-web svg Help Advanced SearchAll fields Title Author Abstract Comments Journal reference ACM classificationMSC classification Report number arXiv identifier DOI ORCID arXiv author IDHelp pages Full textSearchLogin SearchSearch v released -- Feedback?Search term or termsField All fieldsTitleAuthor s AbstractCommentsJournal referenceACMclassificationMSC classificationReport numberarXiv identifierDOIORCIDLicense URI arXiv author IDHelp pagesFull textSearchShow abstracts Hide abstractsAdvanced Search Searching by Author Name Using the Author s field produces best results for author name searches For the most precise name search follow surname s forename s or surname s initial s pattern example Hawking S or Hawking Stephen For best results on multiple author names separate individuals with a semicolon Example Jin D S Ye J Author names enclosed in quotes will return only exact matches For example Stephen Hawking will not return matches for Stephen W Hawking Diacritic character variants are automatically searched in the Author s field Queries with no punctuation will treat each term independently TipsWildcards Use ? to replace a single character or to replace any number of characters Can be used in any field but not in the first character position See Journal References tips for exceptions Expressions TeX expressions can be searched enclosed in single characters Phrases Enclose phrases in double quotes for exact matches in title abstract and comments Dates Sorting by announcement date will use the year and month the _original version_ v of the paper was announced Sorting by submission date will use the year month and day the _latest version_ of the paper was submitted Journal References If a journal reference search contains a wildcard matches will be made using wildcard matching as expected For example math will match _math_ _maths_ _mathematics_ If a journal reference search does not contain a wildcard only exact phrases entered will be matched For example math would match _math_ or _math and science_ but not _maths_ or _mathematics_ All journal reference searches that do not contain a wildcard are literal searches a search for Physica A will match all papers with journal references containing _Physica A_ but a search for Physica A will only return the paper with journal reference _Physica A _ Search v released -- Feedback? About arXiv Leadership Team __Contact __Follow us on Twitter Help Privacy Policy Blog SubscribearXiv® is a registered trademark of Cornell University If you have a disability and are having trouble accessing information on thiswebsite or need materials in an alternate format contact web-accessibility cornell edu for assistance "

Analiza atrybutów dokumentów w programie "Weka"

Analiza z wykorzystaniem filtrów

Poniżej zaprezentowano wyniki w zależności od przyjętych filtrów:

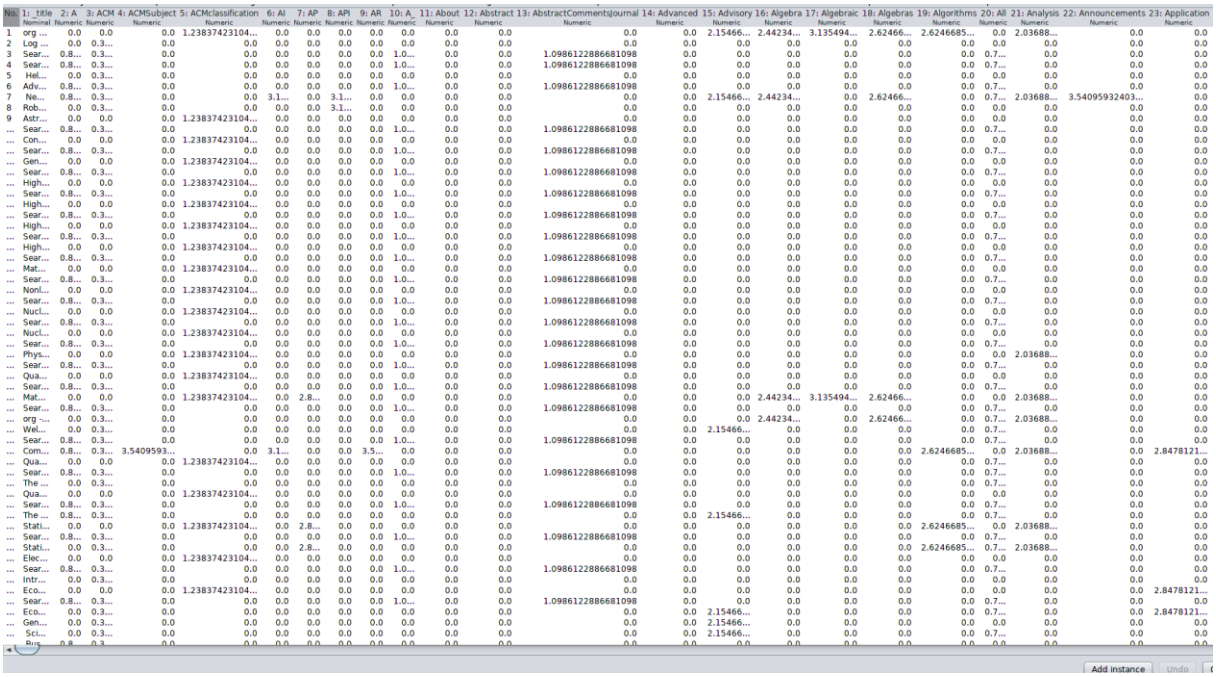
Analiza bez ustawionych parametrów

No.	1: _title	2: A	3: ACM	4: ACMSubject	5: ACMclassification	6: AI	7: AP	8: API	9: AR	10: A_	11: About	12[redacted]					7: Algebraic	18: Algebras	19: Algorithms	20: All	21: Analysis	22: Announcements
	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	org ...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0
2	Log ...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
4	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
5	Hel...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	Adv...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
7	Ne...	1.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0
8	Rob...	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	Astr...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Con...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Gen...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	High...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	High...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	High...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	High...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Mat...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Nonl...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Nucl...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Nucl...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Phys...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Qua...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Mat...	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	org -...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0
...	Wel...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Com...	1.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
...	Qua...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	The ...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Qua...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	The ...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Stati...	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Stati...	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0
...	Elec...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Intr...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Eco...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Eco...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Gen...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sci...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0

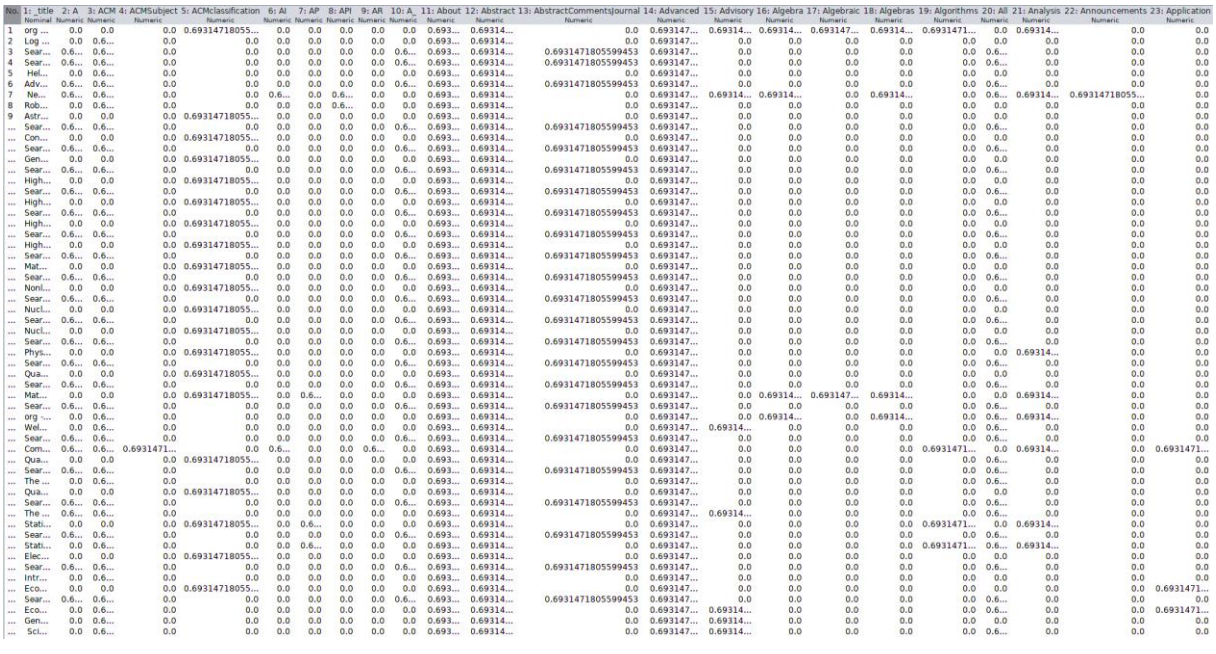
Analiza z ustawionym filtrem NumericToBinary

No.	1: _title	2: A_binarized	3: ACM_binarized	4: ACMSubject_binarized	5: ACMclassification_binarized	6: AI_binarized	7: AP_binarized	8: API_binarized	9: AR_binarized	10: A__binarized	11: About_binarized	12: Abstract_binarized	13: AbstractCommentsJournal_binarized	14: Advanced_binarized	15:
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	
1	org ...	0	0	0	1	0	0	0	0	0	1	1	0	1	1
2	Log ...	0	1	0	0	0	0	0	0	0	1	1	0	1	0
3	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
4	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
5	Hel...	0	1	0	0	0	0	0	0	0	1	1	0	1	0
6	Adv...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
7	Ne...	1	1	0	0	1	0	1	0	0	1	1	0	1	1
8	Rob...	0	1	0	0	0	0	1	0	0	1	1	0	1	0
9	Astr...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	Con...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	Gen...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	High...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	High...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	High...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	High...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	Mat...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	Nonl...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	Nucl...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	Nucl...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	Phys...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	Qua...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	Mat...	0	0	0	1	0	1	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	org -...	0	1	0	0	0	0	0	0	0	1	1	0	1	0
...	Wel...	0	1	0	0	0	0	0	0	0	1	1	0	1	1
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	Com...	1	1	1	0	1	0	0	1	0	1	1	0	1	0
...	Qua...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	The ...	0	1	0	0	0	0	0	0	0	1	1	0	1	0
...	Qua...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	The ...	1	1	0	0	0	0	0	0	0	1	1	0	1	1
...	Stati...	0	0	0	1	0	1	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	Stati...	0	1	0	0	0	1	0	0	0	1	1	0	1	0
...	Elec...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	Intr...	0	1	0	0	0	0	0	0	0	1	1	0	1	0
...	Eco...	0	0	0	1	0	0	0	0	0	1	1	0	1	0
...	Sear...	1	1	0	0	0	0	0	0	1	1	1	1	1	0
...	Eco...	0	1	0	0	0	0	0	0	0	1	1	0	1	1
...	Gen...	0	1	0	0	0	0	0	0	0	1	1	0	1	1
...	Sci...	0	1	0	0	0	0	0	0	0	1	1	0	1	1

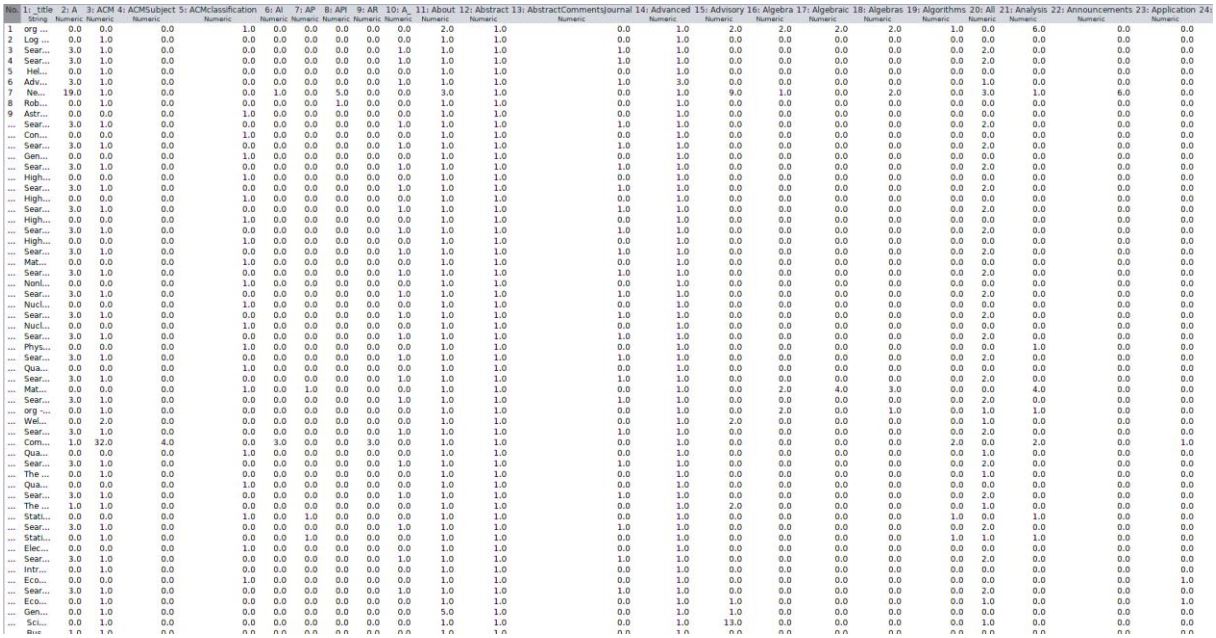
Analiza z ustawionym filtrem IDFTransform



Analiza z ustawionym filtrem TFTransform



Analiza z ustawionym filtrem word-count



No.	1: title String	2: A Numeric	3: ACM Numeric	4: ACMSubject Numeric	5: ACMclassification Numeric	6: AI Numeric	7: AP Numeric	8: API Numeric	9: AR Numeric	10: A_ Numeric	11: About Numeric	12: Abstract Numeric	13: AbstractComments Numeric	Journal 14: Advanced Numeric	15: Advisory Numeric	16: Algebra Numeric	17: Algebraic Numeric	18: Algebras Numeric	19: Algorithms Numeric	20: All Numeric	21: Analysis Numeric	22: Announcements Numeric	23: Application Numeric	24: Numeric
1	org ...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	1.0		0.0	1.0	2.0	2.0	2.0	2.0	1.0	0.0	6.0	0.0	0.0
2	Log ...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
4	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
5	Hel...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	Adv...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	3.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
7	Ne...	19.0	1.0	0.0	0.0	1.0	0.0	5.0	0.0	0.0	3.0	1.0		0.0	1.0	9.0	1.0	0.0	2.0	0.0	3.0	1.0	6.0	0.0
8	Rob...	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	Astr...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	Con...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	Gen...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	High...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	High...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	High...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	Mat...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	Nonl...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	Nucl...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	Nucl...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	Phys...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	Qua...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	Mat...	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	2.0	4.0	3.0	0.0	0.0	4.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	org -...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	2.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0
...	Wei...	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	2.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	Com...	1.0	32.0	4.0	0.0	3.0	0.0	0.0	3.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	2.0	0.0	2.0	0.0	1.0
...	Qua...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	The ...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
...	Qua...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	The ...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	2.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
...	Stati...	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	Stati...	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0
...	Elec...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	Intr...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Eco...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
...	Sear...	3.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0		1.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
...	Eco...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
...	Gen...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	1.0		0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	Sci...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	13.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
...	Rus...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0		0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Podsumowanie

- W analizie bez ustawianych parametrów komórki wypełnione są wartością 1.0 lub 0.0 co oznacza, że dane słowo znajduje sie w dokumencie.
- Przy zastosowaniu filtra NumericToBinary zawartość tablicy jest analogiczna.
- Parametr IDF pozwala na dokonanie transformacji w wyniku którego otrzymuje się macierz z odwrotną częstością dokumentów w dokumencie.
- Parametr TFT dokonuje transformacji TF, która zwraca macierz z wartościami z przedziału <0, 1>, które odpowiadają częstości dokumentów.
- Filtr word-count zwraca macierz, która bezpośrednio odpowiada liczbie słów w dokumencie

Analiza klastrowa

Poniżej znajdują się wyniki klastrowania dla:

K-Means = 3 dla danych bez filtrów:

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: dokumenty-weka.filters.unsupervised.attribute.StringToNominal-Rfirst-weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"

Instances: 69

Attributes: 1116

[list of attributes omitted]

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 6

Within cluster sum of squared errors: 5318.0752380952445

Initial starting points (random):

Cluster 0: {0 'Computer Science Subject Areas and Moderators e-40',1 1,5 1,6 1,7 1,9 1,12 1,14 1,15 1,17 1,22 1,24 1,26 1,31 1,40 1,41 1,42 1,44 1,49 1,58 1,59 1,61 1,62 1,63 1,65 1,68 1,69 1,70 1,71 1,72 1,74 1,75 1,76 1,80 1,83 1,85 1,87 1,94 1,95 1,98 1,104 1,112 1,117 1,121 1,122 1,123 1,124 1,127 1,128 1,133 1,134 1,137 1,138 1,140 1,146 1,148 1,149 1,150 1,151 1,153 1,154 1,156 1,160 1,162 1,167 1,171 1,173 1,175 1,176 1,177 1,179 1,180 1,182 1,186 1,192 1,193 1,200 1,203 1,208 1,209 1,210 1,211 1,214 1,215 1,216 1,221 1,225 1,227 1,232 1,239 1,240 1,241 1,242 1,252 1,254 1,256 1,258 1,267 1,268 1,269 1,272 1,274 1,275 1,277 1,282 1,286 1,289 1,296 1,297 1,298 1,299 1,306 1,307 1,309 1,310 1,316 1,322 1,323 1,326 1,328 1,330 1,333 1,339 1,341 1,345 1,350 1,358 1,371 1,372 1,383 1,387 1,392 1,394 1,396 1,398 1,399 1,403 1,410 1,412 1,414 1,415 1,416 1,418 1,419 1,423 1,431 1,432 1,433 1,434 1,436 1,439 1,440 1,441 1,445 1,446 1,447 1,448 1,449 1,450 1,453 1,455 1,456 1,463 1,464 1,465 1,468 1,469 1,472 1,480 1,481 1,485 1,488 1,489 1,495 1,496 1,497 1,501 1,503 1,504 1,506 1,510 1,512 1,513 1,514 1,515 1,517 1,519 1,523 1,527 1,529 1,530 1,533 1,536 1,538 1,541 1,542 1,552 1,559 1,560 1,561 1,563 1,564 1,565 1,566 1,567 1,574 1,576 1,582 1,583 1,587 1,590 1,593 1,599 1,603 1,608 1,615 1,616 1,622 1,627 1,631 1,634 1,635 1,638 1,639 1,644 1,645 1,646 1,654 1,658 1,659 1,665 1,666 1,672 1,678 1,682 1,685 1,686 1,687 1,688 1,691 1,692 1,693 1,694 1,695 1,699 1,702 1,707 1,708 1,709 1,710 1,711 1,714 1,715 1,721 1,722 1,727 1,728 1,729 1,730 1,731 1,736 1,740 1,741 1,745 1,747 1,748 1,755 1,756 1,759 1,763 1,764 1,765 1,768 1,773 1,777 1,778 1,787 1,794 1,796 1,797 1,798 1,799 1,809 1,810 1,816 1,820 1,824 1,825 1,833 1,834 1,836 1,839 1,843 1,845 1,849 1,851 1,853 1,854 1,857 1,861 1,867 1,874 1,877 1,879 1,882 1,883 1,885 1,887 1,888 1,896 1,904 1,908 1,911 1,912 1,915 1,922 1,925 1,926 1,931 1,933 1,936 1,940 1,942 1,950 1,954 1,956 1,966 1,967 1,969 1,974 1,975 1,976 1,977 1,978 1,982 1,987 1,991 1,993 1,995 1,1000 1,1006 1,1008 1,1011 1,1012 1,1013 1,1015 1,1019 1,1022 1,1023 1,1028 1,1029 1,1030 1,1033 1,1034 1,1035 1,1036 1,1037 1,1038 1,1039 1,1046 1,1047 1,1049 1,1052 1,1054 1,1055 1,1063 1,1065 1,1066 1,1067 1,1069 1,1084 1,1086 1,1087 1,1097 1,1098 1,1100 1,1106 1,1109 1,1114 1}

Cluster 1: {0 'Nonlinear Sciences httpsarxivorgarchivenlin Page 25',1 1,4 1,8 1,14 1,15 1,17 1,29 1,32 1,38 1,40 1,42 1,44 1,46 1,55 1,56 1,57 1,65 1,75 1,83 1,84 1,88 1,96 1,97 1,119 1,123 1,124 1,127 1,132 1,140 1,147 1,151 1,163 1,164 1,167 1,168 1,170 1,178 1,179 1,185 1,188 1,195 1,220 1,222 1,226 1,228 1,230 1,235 1,239 1,242 1,252 1,254 1,267 1,283 1,285 1,292 1,296 1,316 1,323 1,326 1,339 1,345 1,350 1,352 1,358 1,371 1,372 1,380 1,383 1,388 1,394 1,395 1,396 1,412 1,415 1,418 1,419 1,423 1,434 1,435 1,437 1,439 1,446 1,450 1,456 1,470 1,477 1,481 1,482 1,496 1,500 1,513 1,519 1,525 1,529 1,530 1,534 1,537 1,538 1,560 1,563 1,572 1,573 1,583 1,596 1,616 1,631 1,635 1,636 1,638 1,647 1,650 1,652 1,653 1,659 1,665 1,666 1,678 1,680 1,682 1,685 1,690 1,691 1,699 1,701 1,702 1,704 1,705 1,712 1,714 1,717 1,725 1,726 1,727 1,737 1,741 1,743 1,744 1,756 1,761 1,768 1,773 1,778 1,785 1,786 1,788 1,796 1,800 1,805 1,808 1,816 1,820 1,824 1,834 1,836 1,837 1,841 1,845 1,848 1,853 1,857 1,859 1,865 1,878 1,879 1,896 1,904 1,907 1,908 1,911 1,930 1,934 1,938 1,945 1,946 1,951 1,967 1,970 1,972 1,981 1,987 1,988 1,989 1,991 1,993 1,1008 1,1012 1,1015 1,1022 1,1029 1,1039 1,1043 1,1046 1,1048 1,1050 1,1052 1,1054 1,1063 1,1079 1,1086 1,1090 1,1101 1,1106 1,1107 1,1108 1,1112 1,1114 1,1115 1}

Cluster 2: {0 'To submit an article e-print repository63',1 1,2 1,3 1,4 1,6 1,14 1,15 1,17 1,32 1,40 1,44 1,65 1,75 1,76 1,83 1,101 1,103 1,105 1,121 1,124 1,126 1,127 1,140 1,141 1,148 1,151 1,153 1,167 1,174 1,177 1,179 1,206 1,219 1,221 1,227 1,236 1,239 1,252 1,253 1,254 1,263 1,267 1,286 1,290 1,296 1,301 1,312 1,313 1,314 1,316 1,318 1,325 1,326 1,328 1,332 1,333 1,336 1,337 1,339 1,340 1,344 1,345 1,349 1,350 1,354 1,358 1,360 1,362 1,365 1,366 1,371 1,372 1,383 1,384 1,387 1,391 1,394 1,396 1,404 1,406 1,412 1,414 1,415 1,418 1,422 1,423 1,425 1,426 1,427 1,428 1,434 1,437 1,439 1,443 1,444 1,446 1,447 1,448 1,450 1,451 1,453 1,456 1,457 1,460 1,461 1,463 1,464 1,465 1,467 1,469 1,471 1,480 1,481 1,483 1,485 1,486 1,487 1,491 1,492 1,493 1,496 1,497 1,517 1,519 1,529 1,530 1,535 1,538 1,541 1,543 1,544 1,560 1,565 1,566 1,570 1,571 1,574 1,576 1,577 1,583 1,595 1,599 1,610 1,613 1,616 1,618 1,619 1,620 1,630 1,631 1,633 1,635 1,637 1,638 1,643 1,645 1,649 1,659 1,664 1,665 1,666 1,667 1,678 1,682 1,685 1,686 1,688 1,691 1,692 1,695 1,699 1,702 1,703 1,714 1,715 1,716 1,724 1,728 1,732 1,734 1,739 1,742 1,745 1,747 1,749 1,756 1,763 1,768 1,772 1,783 1,788 1,791 1,793 1,795 1,796 1,797 1,800 1,803 1,810 1,811 1,816 1,820 1,824 1,826 1,828 1,829 1,834 1,835 1,836 1,838 1,839 1,841 1,844 1,845 1,857 1,861 1,865 1,869 1,876 1,877 1,879 1,885 1,886 1,888 1,889 1,902 1,903 1,904 1,908 1,910 1,911 1,912 1,920 1,921 1,922 1,925 1,932 1,933 1,938 1,939 1,949 1,950 1,952 1,956 1,966 1,972 1,974 1,979 1,980 1,987 1,996 1,997 1,999 1,1000 1,1006 1,1008 1,1012 1,1014 1,1022 1,1024 1,1028 1,1029 1,1030 1,1031 1,1032 1,1035 1,1036 1,1038 1,1039 1,1042 1,1043 1,1045 1,1046 1,1052 1,1054 1,1057 1,1059 1,1063 1,1065 1,1066 1,1067 1,1068 1,1071 1,1073 1,1075 1,1076 1,1084 1,1086 1,1087 1,1097 1,1098 1,1100 1,1105 1,1106 1,1109 1,1114 1,1115 1}

Missing values globally replaced with mean/mode

Time taken to build model (full training data) : 0.25 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 2 (3%)

1 42 (61%)

2 25 (36%)

K-Means = 5 dla danych bez filtrów:

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: dokumenty-weka.filters.unsupervised.attribute.StringToNominal-Rfirst-weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"

Instances: 69

Attributes: 1116

[list of attributes omitted]

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 3

Within cluster sum of squared errors: 2574.789525691698

Initial starting points (random):

Cluster 0: {0 'Computer Science Subject Areas and Moderators e-40',1 1,5 1,6 1,7 1,9 1,12 1,14 1,15 1,17 1,22 1,24 1,26 1,31 1,40 1,41 1,42 1,44 1,49 1,58 1,59 1,61 1,62 1,63 1,65 1,68 1,69 1,70 1,71 1,72 1,74 1,75 1,76 1,80 1,83 1,85 1,87 1,94 1,95 1,98 1,104 1,112 1,117 1,121 1,122 1,123 1,124 1,127 1,128 1,133 1,134 1,137 1,138 1,140 1,146 1,148 1,149 1,150 1,151 1,153 1,154 1,156 1,160 1,162 1,167 1,171 1,173 1,175 1,176 1,177 1,179 1,180 1,182 1,186 1,192 1,193 1,200 1,203 1,208 1,209 1,210 1,211 1,214 1,215 1,216 1,221 1,225 1,227 1,232 1,239 1,240 1,241 1,242 1,252 1,254 1,256 1,258 1,267 1,268 1,269 1,272 1,274 1,275 1,277 1,282 1,286 1,289 1,296 1,297 1,298 1,299 1,306 1,307 1,309 1,310 1,316 1,322 1,323 1,326 1,328 1,330 1,333 1,339 1,341 1,345 1,350 1,358 1,371 1,372 1,383 1,387 1,392 1,394 1,396 1,398 1,399 1,403 1,410 1,412 1,414 1,415 1,416 1,418 1,419 1,423 1,431 1,432 1,433 1,434 1,436 1,439 1,440 1,441 1,445 1,446 1,447 1,448 1,449 1,450 1,453 1,455 1,456 1,463 1,464 1,465 1,468 1,469 1,472 1,480 1,481 1,485 1,488 1,489 1,495 1,496 1,497 1,501 1,503 1,504 1,506 1,510 1,512 1,513 1,514 1,515 1,517 1,519 1,523 1,527 1,529 1,530 1,533 1,536 1,538 1,541 1,542 1,552 1,559 1,560 1,561 1,563 1,564 1,565 1,566 1,567 1,574 1,576 1,582 1,583 1,587 1,590 1,593 1,599 1,603 1,608 1,615 1,616 1,622 1,627 1,631 1,634 1,635 1,638 1,639 1,644 1,645 1,646 1,654 1,658 1,659 1,665 1,666 1,672 1,678 1,682 1,685 1,686 1,687 1,688 1,691 1,692 1,693 1,694 1,695 1,699 1,702 1,707 1,708 1,709 1,710 1,711 1,714 1,715 1,721 1,722 1,727 1,728 1,729 1,730 1,731 1,736 1,740 1,741 1,745 1,747 1,748 1,755 1,756 1,759 1,763 1,764 1,765 1,768 1,773 1,777 1,778 1,787 1,794 1,796 1,797 1,798 1,799 1,809 1,810 1,816 1,820 1,824 1,825 1,833 1,834 1,836 1,839 1,843 1,845 1,849 1,851 1,853 1,854 1,857 1,861 1,867 1,874 1,877 1,879 1,882 1,883 1,885 1,887 1,888 1,896 1,904 1,908 1,911 1,912 1,915 1,922 1,925 1,926 1,931 1,933 1,936 1,940 1,942 1,950 1,954 1,956 1,966 1,967 1,969 1,974 1,975 1,976 1,977 1,978 1,982 1,987 1,991 1,993 1,995 1,1000 1,1006 1,1008 1,1011 1,1012 1,1013 1,1015 1,1019 1,1022 1,1023 1,1028 1,1029 1,1030 1,1033 1,1034 1,1035 1,1036 1,1037 1,1038 1,1039 1,1046 1,1047 1,1049 1,1052 1,1054 1,1055 1,1063 1,1065 1,1066 1,1067 1,1069 1,1084 1,1086 1,1087 1,1097 1,1098 1,1100 1,1106 1,1109 1,1114 1}

Cluster 1: {0 'Nonlinear Sciences httpsarxivorgarchivenlin Page 25',1 1,4 1,8 1,14 1,15 1,17 1,29 1,32 1,38 1,40 1,42 1,44 1,46 1,55 1,56 1,57 1,65 1,75 1,83 1,84 1,88 1,96 1,97 1,119 1,123 1,124 1,127 1,132 1,140 1,147 1,151 1,163 1,164 1,167 1,168 1,170 1,178 1,179 1,185 1,188 1,195 1,220 1,222 1,226 1,228 1,230 1,235 1,239 1,242 1,252 1,254 1,267 1,283 1,285 1,292 1,296 1,316 1,323 1,326 1,339 1,345 1,350 1,352 1,358 1,371 1,372 1,380 1,383 1,388 1,394 1,395 1,396 1,412 1,415 1,418 1,419 1,423 1,434 1,435 1,437 1,439 1,446 1,450 1,456 1,470 1,477 1,481 1,482 1,496 1,500 1,513 1,519 1,525 1,529 1,530 1,534 1,537 1,538 1,560 1,563 1,572 1,573 1,583 1,596 1,616 1,631 1,635 1,636 1,638 1,647 1,650 1,652 1,653 1,659 1,665 1,666 1,678 1,680 1,682 1,685 1,690 1,691 1,699 1,701 1,702 1,704 1,705 1,712 1,714 1,717 1,725 1,726 1,727 1,737 1,741 1,743 1,744 1,756 1,761 1,768 1,773 1,778 1,785 1,786 1,788 1,796 1,800 1,805 1,808 1,816 1,820 1,824 1,834 1,836 1,837 1,841 1,845 1,848 1,853 1,857 1,859 1,865 1,878 1,879 1,896 1,904 1,907 1,908 1,911 1,930 1,934 1,938 1,945 1,946 1,951 1,967 1,970 1,972 1,981 1,987 1,988 1,989 1,991 1,993 1,1008 1,1012 1,1015 1,1022 1,1029 1,1039 1,1043 1,1046 1,1048 1,1050 1,1052 1,1054 1,1063 1,1079 1,1086 1,1090 1,1101 1,1106 1,1107 1,1108 1,1112 1,1114 1,1115 1}

Cluster 2: {0 'To submit an article e-print repository63',1 1,2 1,3 1,4 1,6 1,14 1,15 1,17 1,32 1,40 1,44 1,65 1,75 1,76 1,83 1,101 1,103 1,105 1,121 1,124 1,126 1,127 1,140 1,141 1,148 1,151 1,153 1,167 1,174 1,177 1,179 1,206 1,219 1,221 1,227 1,236 1,239 1,252 1,253 1,254 1,263 1,267 1,286 1,290 1,296 1,301 1,312 1,313 1,314 1,316 1,318 1,325 1,326 1,328 1,332 1,333 1,336 1,337 1,339 1,340 1,344 1,345 1,349 1,350 1,354 1,358 1,360 1,362 1,365 1,366 1,371 1,372 1,383 1,384 1,387 1,391 1,394 1,396 1,404 1,406 1,412 1,414 1,415 1,418 1,422 1,423 1,425 1,426 1,427 1,428 1,434 1,437 1,439 1,443 1,444 1,446 1,447 1,448 1,450 1,451 1,453 1,456 1,457 1,460 1,461 1,463 1,464 1,465 1,467 1,469 1,471 1,480 1,481 1,483 1,485 1,486 1,487 1,491 1,492 1,493 1,496 1,497 1,517 1,519 1,529 1,530 1,535 1,538 1,541 1,543 1,544 1,560 1,565 1,566 1,570 1,571 1,574 1,576 1,577 1,583 1,595 1,599 1,610 1,613 1,616 1,618 1,619 1,620 1,630 1,631 1,633 1,635 1,637 1,638 1,643 1,645 1,649 1,659 1,664 1,665 1,666 1,667 1,678 1,682 1,685 1,686 1,688 1,691 1,692 1,695 1,699 1,702 1,703 1,714 1,715 1,716 1,724 1,728 1,732 1,734 1,739 1,742 1,745 1,747 1,749 1,756 1,763 1,768 1,772 1,783 1,788 1,791 1,793 1,795 1,796 1,797 1,800 1,803 1,810 1,811 1,816 1,820 1,824 1,826 1,828 1,829 1,834 1,835 1,836 1,838 1,839 1,841 1,844 1,845 1,857 1,861 1,865 1,869 1,876 1,877 1,879 1,885 1,886 1,888 1,889 1,902 1,903 1,904 1,908 1,910 1,911 1,912 1,920 1,921 1,922 1,925 1,932 1,933 1,938 1,939 1,949 1,950 1,952 1,956 1,966 1,972 1,974 1,979 1,980 1,987 1,996 1,997 1,999 1,1000 1,1006 1,1008 1,1012 1,1014 1,1022 1,1024 1,1028 1,1029 1,1030 1,1031 1,1032 1,1035 1,1036 1,1038 1,1039 1,1042 1,1043 1,1045 1,1046 1,1052 1,1054 1,1057 1,1059 1,1063 1,1065 1,1066 1,1067 1,1068 1,1071 1,1073 1,1075 1,1076 1,1084 1,1086 1,1087 1,1097 1,1098 1,1100 1,1105 1,1106 1,1109 1,1114 1,1115 1}

Cluster 3: {0 'Contacting e-print repository httpsarxivorghelp68',2 1,6 1,14 1,15 1,17 1,40 1,44 1,65 1,75 1,76 1,83 1,101 1,103 1,118 1,121 1,124 1,126 1,127 1,140 1,141 1,148 1,151 1,167 1,177 1,179 1,198 1,206 1,221 1,227 1,239 1,251 1,252 1,254 1,267 1,286 1,291 1,296 1,316 1,326 1,339 1,340 1,345 1,350 1,358 1,371 1,372 1,383 1,384 1,394 1,396 1,403 1,412 1,415 1,418 1,423 1,434 1,439 1,446 1,447 1,448 1,450 1,453 1,456 1,458 1,463 1,465 1,483 1,486 1,496 1,497 1,519 1,526 1,529 1,530 1,550 1,554 1,560 1,574 1,576 1,583 1,588 1,599 1,616 1,621 1,631 1,635 1,638 1,643 1,645 1,659 1,665 1,666 1,667 1,678 1,681 1,682 1,685 1,688 1,691 1,695 1,699 1,702 1,703 1,714 1,728 1,756 1,763 1,768 1,770 1,771 1,780 1,783 1,796 1,808 1,816 1,820 1,824 1,834

1,836 1,839 1,841 1,845 1,847 1,857 1,861 1,864 1,873 1,898 1,902 1,904 1,905 1,908 1,911 1,915 1,922 1,924 1,928 1,933 1,938 1,977 1,987 1,995 1,996 1,1008 1,1012 1,1016 1,1024 1,1028 1,1029 1,1035 1,1039 1,1042 1,1043 1,1046 1,1052 1,1054 1,1063 1,1065 1,1086 1,1087 1,1106 1,1114 1,1115 1}

Cluster 4: {0 ' Scientific Advisory Board e-print repository57',2 1,6 1,14 1,15 1,17 1,18 1,23 1,28 1,35 1,36 1,40 1,43 1,44 1,45 1,47 1,52 1,63 1,64 1,65 1,71 1,72 1,75 1,76 1,79 1,83 1,87 1,93 1,121 1,124 1,127 1,140 1,141 1,148 1,151 1,156 1,158 1,167 1,177 1,179 1,193 1,197 1,217 1,227 1,239 1,248 1,249 1,252 1,254 1,257 1,259 1,264 1,267 1,269 1,282 1,283 1,284 1,286 1,296 1,298 1,302 1,309 1,316 1,326 1,328 1,329 1,339 1,345 1,350 1,358 1,371 1,372 1,373 1,383 1,394 1,396 1,415 1,418 1,423 1,434 1,438 1,439 1,446 1,447 1,448 1,450 1,456 1,463 1,474 1,481 1,490 1,496 1,497 1,514 1,519 1,529 1,530 1,532 1,538 1,545 1,560 1,576 1,577 1,582 1,583 1,616 1,622 1,630 1,631 1,635 1,638 1,656 1,659 1,665 1,666 1,667 1,678 1,682 1,685 1,688 1,691 1,699 1,702 1,714 1,716 1,728 1,730 1,756 1,760 1,768 1,769 1,780 1,783 1,796 1,816 1,820 1,822 1,824 1,830 1,834 1,836 1,845 1,851 1,857 1,859 1,861 1,866 1,887 1,895 1,904 1,908 1,911 1,922 1,927 1,933 1,938 1,942 1,957 1,987 1,989 1,1008 1,1012 1,1014 1,1024 1,1028 1,1029 1,1039 1,1046 1,1052 1,1054 1,1063 1,1086 1,1087 1,1097 1,1106 1,1110 1,1114 1}

Missing values globally replaced with mean/mode

Time taken to build model (full training data) : 0.15 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 2 (3%)

1 20 (29%)

2 2 (3%)

3 23 (33%)

4 22 (32%)

K-Means = 3 dla danych z filtrem NumericToBinary:

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: dokumenty-weka.filters.unsupervised.attribute.StringToNominal-Rfirst-weka.filters.unsupervised.attribute.NumericToBinary-Rlast-weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"-weka.filters.unsupervised.attribute.Remove-R2-5-weka.filters.unsupervised.attribute.NumericToBinary-Rfirst-last

Instances: 69

Attributes: 1112

[list of attributes omitted]

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 2

Within cluster sum of squared errors: 8473.0

Initial starting points (random):

Cluster 0: {0 'Computer Science Subject Areas and Moderators e-40',1 1,2 1,3 1,5 1,8 1,10 1,11 1,13 1,18 1,20 1,22 1,27 1,36 1,37 1,38 1,40 1,45 1,54 1,55 1,57 1,58 1,59 1,61 1,64 1,65 1,66 1,67 1,68 1,70 1,71 1,72 1,76 1,79 1,81 1,83 1,90 1,91 1,94 1,100 1,108 1,113 1,117 1,118 1,119 1,120 1,123 1,124 1,129 1,130 1,133 1,134 1,136 1,142 1,144 1,145 1,146 1,147 1,149 1,150 1,152 1,156 1,158 1,163 1,167 1,169 1,171 1,172 1,173 1,175 1,176 1,178 1,182 1,188 1,189 1,196 1,199 1,204 1,205 1,206 1,207 1,210 1,211 1,212 1,217 1,221 1,223 1,228 1,235 1,236 1,237 1,238 1,248 1,250 1,252 1,254 1,263 1,264 1,265 1,268 1,270 1,271 1,273 1,278 1,282 1,285 1,292 1,293 1,294 1,295 1,302 1,303 1,305 1,306 1,312 1,318 1,319 1,322 1,324 1,326 1,329 1,335 1,337 1,341 1,346 1,354 1,367 1,368 1,379 1,383 1,388 1,390 1,392 1,394 1,395 1,399 1,406 1,408 1,410 1,411 1,412 1,414 1,415 1,419 1,427 1,428 1,429 1,430 1,432 1,435 1,436 1,437 1,441 1,442 1,443 1,444 1,445 1,446 1,449 1,451 1,452 1,459 1,460 1,461 1,464 1,465 1,468 1,476 1,477 1,481 1,484 1,485 1,491 1,492 1,493 1,497 1,499 1,500 1,502 1,506 1,508 1,509 1,510 1,511 1,513 1,515 1,519 1,523 1,525 1,526 1,529 1,532 1,534 1,537 1,538 1,548 1,555 1,556 1,557 1,559

1,560 1,561 1,562 1,563 1,570 1,572 1,578 1,579 1,583 1,586 1,589 1,595 1,599 1,604 1,611 1,612 1,618 1,623 1,627 1,630 1,631 1,634 1,635 1,640 1,641 1,642 1,650 1,654 1,655 1,661 1,662 1,668 1,674 1,678 1,681 1,682 1,683 1,684 1,687 1,688 1,689 1,690 1,691 1,695 1,698 1,703 1,704 1,705 1,706 1,707 1,710 1,711 1,717 1,718 1,723 1,724 1,725 1,726 1,727 1,732 1,736 1,737 1,741 1,743 1,744 1,751 1,752 1,755 1,759 1,760 1,761 1,764 1,769 1,773 1,774 1,783 1,790 1,792 1,793 1,794 1,795 1,805 1,806 1,812 1,816 1,820 1,821 1,829 1,830 1,832 1,835 1,839 1,841 1,845 1,847 1,849 1,850 1,853 1,857 1,863 1,870 1,873 1,875 1,878 1,879 1,881 1,883 1,884 1,892 1,900 1,904 1,907 1,908 1,911 1,918 1,921 1,922 1,927 1,929 1,932 1,936 1,938 1,946 1,950 1,952 1,962 1,963 1,965 1,970 1,971 1,972 1,973 1,974 1,978 1,983 1,987 1,989 1,991 1,996 1,1002 1,1004 1,1007 1,1008 1,1009 1,1011 1,1015 1,1018 1,1019 1,1024 1,1025 1,1026 1,1029 1,1030 1,1031 1,1032 1,1033 1,1034 1,1035 1,1042 1,1043 1,1045 1,1048 1,1050 1,1051 1,1059 1,1061 1,1062 1,1063 1,1065 1,1080 1,1082 1,1083 1,1093 1,1094 1,1096 1,1102 1,1105 1,1110 1}

Cluster 1: {0 'Nonlinear Sciences httpsarxivorgarchivenlin Page 25',4 1,10 1,11 1,13 1,25 1,28 1,34 1,36 1,38 1,40 1,42 1,51 1,52 1,53 1,61 1,71 1,79 1,80 1,84 1,92 1,93 1,115 1,119 1,120 1,123 1,128 1,136 1,143 1,147 1,159 1,160 1,163 1,164 1,166 1,174 1,175 1,181 1,184 1,191 1,216 1,218 1,222 1,224 1,226 1,231 1,235 1,238 1,248 1,250 1,263 1,279 1,281 1,288 1,292 1,312 1,319 1,322 1,335 1,341 1,346 1,348 1,354 1,367 1,368 1,376 1,379 1,384 1,390 1,391 1,392 1,408 1,411 1,414 1,415 1,419 1,430 1,431 1,433 1,435 1,442 1,446 1,452 1,466 1,473 1,477 1,478 1,492 1,496 1,509 1,515 1,521 1,525 1,526 1,530 1,533 1,534 1,556 1,559 1,568 1,569 1,579 1,592 1,612 1,627 1,631 1,632 1,634 1,643 1,646 1,648 1,649 1,655 1,661 1,662 1,674 1,676 1,678 1,681 1,686 1,687 1,695 1,697 1,698 1,700 1,701 1,708 1,710 1,713 1,721 1,722 1,723 1,733 1,737 1,739 1,740 1,752 1,757 1,764 1,769 1,774 1,781 1,782 1,784 1,792 1,796 1,801 1,804 1,812 1,816 1,820 1,830 1,832 1,833 1,837 1,841 1,844 1,849 1,853 1,855 1,861 1,874 1,875 1,892 1,900 1,903 1,904 1,907 1,926 1,930 1,934 1,941 1,942 1,947 1,963 1,966 1,968 1,977 1,983 1,984 1,985 1,987 1,989 1,1004 1,1008 1,1011 1,1018 1,1025 1,1035 1,1039 1,1042 1,1044 1,1046 1,1048 1,1050 1,1059 1,1075 1,1082 1,1086 1,1097 1,1102 1,1103 1,1104 1,1108 1,1110 1,1111 1}

Cluster 2: {0 'To submit an article e-print repository63',2 1,10 1,11 1,13 1,28 1,36 1,40 1,61 1,71 1,72 1,79 1,97 1,99 1,101 1,117 1,120 1,122 1,123 1,136 1,137 1,144 1,147 1,149 1,163 1,170 1,173 1,175 1,202 1,215 1,217 1,223 1,232 1,235 1,248 1,249 1,250 1,259 1,263 1,282 1,286 1,292 1,297 1,308 1,309 1,310 1,312 1,314 1,321 1,322 1,324 1,328 1,329 1,332 1,333 1,335 1,336 1,340 1,341 1,345 1,346 1,350 1,354 1,356 1,358 1,361 1,362 1,367 1,368 1,379 1,380 1,383 1,387 1,390 1,392 1,400 1,402 1,408 1,410 1,411 1,414 1,418 1,419 1,421 1,422 1,423 1,424 1,430 1,433 1,435 1,439 1,440 1,442 1,443 1,444 1,446 1,447 1,449 1,452 1,453 1,456 1,457 1,459 1,460 1,461 1,463 1,465 1,467 1,476 1,477 1,479 1,481 1,482 1,483 1,487 1,488 1,489 1,492 1,493 1,513 1,515 1,525 1,526 1,531 1,534 1,537 1,539 1,540 1,556 1,561 1,562 1,566 1,567 1,570 1,572 1,573 1,579 1,591 1,595 1,606 1,609 1,612 1,614 1,615 1,616 1,626 1,627 1,629 1,631 1,633 1,634 1,639 1,641 1,645 1,655 1,660 1,661 1,662 1,663 1,674 1,678 1,681 1,682 1,684 1,687 1,688 1,691 1,695 1,698 1,699 1,710 1,711 1,712 1,720 1,724 1,728 1,730 1,735 1,738 1,741 1,743 1,745 1,752 1,759 1,764 1,768 1,779 1,784 1,787 1,789 1,791 1,792 1,793 1,796 1,799 1,806 1,807 1,812 1,816 1,820 1,822 1,824 1,825 1,830 1,831 1,832 1,834 1,835 1,837 1,840 1,841 1,853 1,857 1,861 1,865 1,872 1,873 1,875 1,881 1,882 1,884 1,885 1,898 1,899 1,900 1,904 1,906 1,907 1,908 1,916 1,917 1,918 1,921 1,928 1,929 1,934 1,935 1,945 1,946 1,948 1,952 1,962 1,968 1,970 1,975 1,976 1,983 1,992 1,993 1,995 1,996 1,1002 1,1004 1,1008 1,1010 1,1018 1,1020 1,1024 1,1025 1,1026 1,1027 1,1028 1,1031 1,1032 1,1034 1,1035 1,1038 1,1039 1,1041 1,1042 1,1048 1,1050 1,1053 1,1055 1,1059 1,1061 1,1062 1,1063 1,1064 1,1067 1,1069 1,1071 1,1072 1,1080 1,1082 1,1083 1,1093 1,1094 1,1096 1,1101 1,1102 1,1105 1,1110 1,1111 1}

Missing values globally replaced with mean/mode

Time taken to build model (full training data) : 0.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 2 (3%)

1 59 (86%)

2 8 (12%)

K-Means = 5 dla danych z filtrem NumericToBinary:

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: dokumenty-weka.filters.unsupervised.attribute.StringToNominal-Rfirst-weka.filters.unsupervised.attribute.NumericToBinary-Rlast-weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"-weka.filters.unsupervised.attribute.Remove-R2-5-weka.filters.unsupervised.attribute.NumericToBinary-Rfirst-last

Instances: 69

Attributes: 1112

[list of attributes omitted]

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 3

Within cluster sum of squared errors: 3290.0

Initial starting points (random):

Cluster 0: {0 'Computer Science Subject Areas and Moderators e-40',1 1,2 1,3 1,5 1,8 1,10 1,11 1,13 1,18 1,20 1,22 1,27 1,36 1,37 1,38 1,40 1,45 1,54 1,55 1,57 1,58 1,59 1,61 1,64 1,65 1,66 1,67 1,68 1,70 1,71 1,72 1,76 1,79 1,81 1,83 1,90 1,91 1,94 1,100 1,108 1,113 1,117 1,118 1,119 1,120 1,123 1,124 1,129 1,130 1,133 1,134 1,136 1,142 1,144 1,145 1,146 1,147 1,149 1,150 1,152 1,156 1,158 1,163 1,167 1,169 1,171 1,172 1,173 1,175 1,176 1,178 1,182 1,188 1,189 1,196 1,199 1,204 1,205 1,206 1,207 1,210 1,211 1,212 1,217 1,221 1,223 1,228 1,235 1,236 1,237 1,238 1,248 1,250 1,252 1,254 1,263 1,264 1,265 1,268 1,270 1,271 1,273 1,278 1,282 1,285 1,292 1,293 1,294 1,295 1,302 1,303 1,305 1,306 1,312 1,318 1,319 1,322 1,324 1,326 1,329 1,335 1,337 1,341 1,346 1,354 1,367 1,368 1,379 1,383 1,388 1,390 1,392 1,394 1,395 1,399 1,406 1,408 1,410 1,411 1,412 1,414 1,415 1,419 1,427 1,428 1,429 1,430 1,432 1,435 1,436 1,437 1,441 1,442 1,443 1,444 1,445 1,446 1,449 1,451 1,452 1,459 1,460 1,461 1,464 1,465 1,468 1,476 1,477 1,481 1,484 1,485 1,491 1,492 1,493 1,497 1,499 1,500 1,502 1,506 1,508 1,509 1,510 1,511 1,513 1,515 1,519 1,523 1,525 1,526 1,529 1,532 1,534 1,537 1,538 1,548 1,555 1,556 1,557 1,559 1,560 1,561 1,562 1,563 1,570 1,572 1,578 1,579 1,583 1,586 1,589 1,595 1,599 1,604 1,611 1,612 1,618 1,623 1,627 1,630 1,631 1,634 1,635 1,640 1,641 1,642 1,650 1,654 1,655 1,661 1,662 1,668 1,674 1,678 1,681 1,682 1,683 1,684 1,687 1,688 1,689 1,690 1,691 1,695 1,698 1,703 1,704 1,705 1,706 1,707 1,710 1,711 1,717 1,718 1,723 1,724 1,725 1,726 1,727 1,732 1,736 1,737 1,741 1,743 1,744 1,751 1,752 1,755 1,759 1,760 1,761 1,764 1,769 1,773 1,774 1,783 1,790 1,792 1,793 1,794 1,795 1,805 1,806 1,812 1,816 1,820 1,821 1,829 1,830 1,832 1,835 1,839 1,841 1,845 1,847 1,849 1,850 1,853 1,857 1,863 1,870 1,873 1,875 1,878 1,879 1,881 1,883 1,884 1,892 1,900 1,904 1,907 1,908 1,911 1,918 1,921 1,922 1,927 1,929 1,932 1,936 1,938 1,946 1,950 1,952 1,962 1,963 1,965 1,970 1,971 1,972 1,973 1,974 1,978 1,983 1,987 1,989 1,991 1,996 1,1002 1,1004 1,1007 1,1008 1,1009 1,1011 1,1015 1,1018 1,1019 1,1024 1,1025 1,1026 1,1029 1,1030 1,1031 1,1032 1,1033 1,1034 1,1035 1,1042 1,1043 1,1045 1,1048 1,1050 1,1051 1,1059 1,1061 1,1062 1,1063 1,1065 1,1080 1,1082 1,1083 1,1093 1,1094 1,1096 1,1102 1,1105 1,1110 1}

Cluster 1: {0 'Nonlinear Sciences httpsarxivorgarchivenlin Page 25',4 1,10 1,11 1,13 1,25 1,28 1,34 1,36 1,38 1,40 1,42 1,51 1,52 1,53 1,61 1,71 1,79 1,80 1,84 1,92 1,93 1,115 1,119 1,120 1,123 1,128 1,136 1,143 1,147 1,159 1,160 1,163 1,164 1,166 1,174 1,175 1,181 1,184 1,191 1,216 1,218 1,222 1,224 1,226 1,231 1,235 1,238 1,248 1,250 1,263 1,279 1,281 1,288 1,292 1,312 1,319 1,322 1,335 1,341 1,346 1,348 1,354 1,367 1,368 1,376 1,379 1,384 1,390 1,391 1,392 1,408 1,411 1,414 1,415 1,419 1,430 1,431 1,433 1,435 1,442 1,446 1,452 1,466 1,473 1,477 1,478 1,492 1,496 1,509 1,515 1,521 1,525 1,526 1,530 1,533 1,534 1,556 1,559 1,568 1,569 1,579 1,592 1,612 1,627 1,631 1,632 1,634 1,643 1,646 1,648 1,649 1,655 1,661 1,662 1,674 1,676 1,678 1,681 1,686 1,687 1,695 1,697 1,698 1,700 1,701 1,708 1,710 1,713 1,721 1,722 1,723 1,733 1,737 1,739 1,740 1,752 1,757 1,764 1,769 1,774 1,781 1,782 1,784 1,792 1,796 1,801 1,804 1,812 1,816 1,820 1,830 1,832 1,833 1,837 1,841 1,844 1,849 1,853 1,855 1,861 1,874 1,875 1,892 1,900 1,903 1,904 1,907 1,926 1,930 1,934 1,941 1,942 1,947 1,963 1,966 1,968 1,977 1,983 1,984 1,985 1,987 1,989 1,1004 1,1008 1,1011 1,1018 1,1025 1,1035 1,1039 1,1042 1,1044 1,1046 1,1048 1,1050 1,1059 1,1075 1,1082 1,1086 1,1097 1,1102 1,1103 1,1104 1,1108 1,1110 1,1111 1}

Cluster 2: {0 'To submit an article e-print repository63',2 1,10 1,11 1,13 1,28 1,36 1,40 1,61 1,71 1,72 1,79 1,97 1,99 1,101 1,117 1,120 1,122 1,123 1,136 1,137 1,144 1,147 1,149 1,163 1,170 1,173 1,175 1,202 1,215 1,217 1,223 1,232 1,235 1,248 1,249 1,250 1,259 1,263 1,282 1,286 1,292 1,297 1,308 1,309 1,310 1,312 1,314 1,321 1,322 1,324 1,328 1,329 1,332 1,333 1,335 1,336 1,340 1,341 1,345 1,346 1,350 1,354 1,356 1,358 1,361 1,362 1,367 1,368 1,379 1,380 1,383 1,387 1,390 1,392 1,400 1,402 1,408 1,410 1,411 1,414 1,418 1,419 1,421 1,422 1,423 1,424 1,430 1,433 1,435 1,439 1,440 1,442 1,443 1,444 1,446 1,447 1,449 1,452 1,453 1,456 1,457 1,459 1,460 1,461 1,463 1,465 1,467 1,476 1,477 1,479 1,481 1,482 1,483 1,487 1,488 1,489 1,492 1,493 1,513 1,515 1,525 1,526 1,531 1,534 1,537 1,539 1,540 1,556 1,561 1,562 1,566 1,567 1,570 1,572 1,573 1,579 1,591 1,595 1,606 1,609 1,612 1,614 1,615 1,616 1,626 1,627 1,629 1,631 1,633 1,634 1,639 1,641 1,645 1,655 1,660 1,661 1,662 1,663 1,674 1,678 1,681 1,682 1,684 1,687 1,688 1,691 1,695 1,698 1,699 1,710 1,711 1,712 1,720 1,724 1,728 1,730 1,735 1,738 1,741 1,743 1,745 1,752 1,759 1,764 1,768 1,779 1,784 1,787 1,789 1,791 1,792 1,793 1,796 1,799 1,806 1,807 1,812 1,816 1,820 1,822 1,824 1,825 1,830 1,831 1,832 1,834 1,835 1,837 1,840 1,841 1,853 1,857 1,861 1,865 1,872 1,873 1,875 1,881 1,882 1,884 1,885 1,898 1,899 1,900 1,904 1,906 1,907 1,908 1,916 1,917 1,918 1,921 1,928 1,929 1,934 1,935 1,945 1,946 1,948 1,952 1,962 1,968 1,970 1,975 1,976 1,983 1,992 1,993 1,995 1,996 1,1002 1,1004 1,1008 1,1010 1,1018 1,1020 1,1024 1,1025 1,1026 1,1027 1,1028 1,1031 1,1032 1,1034 1,1035 1,1038 1,1039 1,1041 1,1042 1,1048 1,1050 1,1053 1,1055 1,1059 1,1061 1,1062 1,1063 1,1064 1,1067 1,1069 1,1071 1,1072 1,1080 1,1082 1,1083 1,1093 1,1094 1,1096 1,1101 1,1102 1,1105 1,1110 1,1111 1}

Cluster 3: {0 'Contacting e-print repository httpsarxivorghelp68',2 1,10 1,11 1,13 1,36 1,40 1,61 1,71 1,72 1,79 1,97 1,99 1,114 1,117 1,120 1,122 1,123 1,136 1,137 1,144 1,147 1,163 1,173 1,175 1,194 1,202 1,217 1,223 1,235 1,247 1,248 1,250 1,263 1,282 1,287 1,292 1,312 1,322 1,335 1,336 1,341 1,346 1,354 1,367 1,368 1,379 1,380 1,390 1,392 1,399 1,408 1,411 1,414 1,419 1,430 1,435 1,442 1,443 1,444 1,446 1,449 1,452 1,454 1,459 1,461 1,479 1,482 1,492 1,493 1,515 1,522 1,525 1,526 1,546 1,550 1,556 1,570 1,572 1,579 1,584 1,595 1,612 1,617 1,627 1,631 1,634 1,639 1,641 1,655 1,661 1,662 1,663 1,674 1,677 1,678 1,681 1,684 1,687 1,691 1,695 1,698 1,699 1,710 1,724 1,752 1,759 1,764 1,766 1,767 1,776 1,779 1,792 1,804 1,812 1,816 1,820 1,830 1,832 1,835 1,837 1,841 1,843 1,853 1,857 1,860 1,869 1,894 1,898 1,900 1,901 1,904 1,907 1,911 1,918 1,920 1,924 1,929 1,934 1,973 1,983 1,991 1,992 1,1004 1,1008 1,1012 1,1020 1,1024 1,1025 1,1031 1,1035 1,1038 1,1039 1,1042 1,1048 1,1050 1,1059 1,1061 1,1082 1,1083 1,1102 1,1110 1,1111 1}

Cluster 4: {0 ' Scientific Advisory Board e-print repository57',2 1,10 1,11 1,13 1,14 1,19 1,24 1,31 1,32 1,36 1,39 1,40 1,41 1,43 1,48 1,59 1,60 1,61 1,67 1,68 1,71 1,72 1,75 1,79 1,83 1,89 1,117 1,120 1,123 1,136 1,137 1,144 1,147 1,152 1,154 1,163 1,173 1,175 1,189 1,193 1,213 1,223 1,235 1,244 1,245 1,248 1,250 1,253 1,255 1,260 1,263 1,265 1,278 1,279 1,280 1,282 1,292 1,294 1,298 1,305 1,312 1,322 1,324 1,325 1,335 1,341 1,346 1,354 1,367 1,368 1,369 1,379 1,390 1,392 1,411 1,414 1,419 1,430 1,434 1,435 1,442 1,443 1,444 1,446 1,452 1,459 1,470 1,477 1,486 1,492 1,493 1,510 1,515 1,525 1,526 1,528 1,534 1,541 1,556 1,572 1,573 1,578 1,579 1,612 1,618 1,626 1,627 1,631 1,634 1,652 1,655 1,661 1,662 1,663 1,674 1,678 1,681 1,684 1,687 1,695 1,698 1,710 1,712 1,724 1,726 1,752 1,756 1,764 1,765 1,776 1,779 1,792 1,812 1,816 1,818 1,820 1,826 1,830 1,832 1,841 1,847 1,853 1,855 1,857 1,862 1,883 1,891 1,900 1,904 1,907 1,918 1,923 1,929 1,934 1,938 1,953 1,983 1,985 1,1004 1,1008 1,1010 1,1020 1,1024 1,1025 1,1035 1,1042 1,1048 1,1050 1,1059 1,1082 1,1083 1,1093 1,1102 1,1106 1,1110 1}

Missing values globally replaced with mean/mode

Time taken to build model (full training data) : 0.06 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 2 (3%)

1 20 (29%)

2 3 (4%)

3 23 (33%)

4 21 (30%)

K-Means = 3 dla danych z filtrem IDFTransform:

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: dokumenty-weka.filters.unsupervised.attribute.StringToNominal-Rfirst-weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-I-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"-weka.filters.unsupervised.attribute.Remove-R2-5

Instances: 69

Attributes: 1112

[list of attributes omitted]

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 6

Within cluster sum of squared errors: 5275.052380952385

Initial starting points (random):

Cluster 0: {0 'Computer Science Subject Areas and Moderators e-40',1 0.832909,2 0.342286,3 3.540959,5 3.135494,8 3.540959,18 2.624669,20 2.036882,22 2.847812,27 3.135494,37 2.624669,38 2.847812,45 2.847812,54 3.135494,55 3.135494,57 2.442347,58 3.135494,59 2.624669,64 3.135494,65 2.288196,66 2.442347,67 1.931521,68 2.442347,70 2.288196,72 0.362905,76 0.97601,81 2.154665,83 2.624669,90 2.847812,91 3.135494,94 3.540959,100 3.135494,108 2.036882,113 3.540959,117 0.362905,118 2.154665,119 0.342286,124 3.540959,129 1.526056,130 2.847812,133 3.135494,134 2.847812,142 2.847812,144 0.342286,145 3.135494,146 2.624669,149 1.931521,150 2.624669,152 1.7492,156 3.135494,158 1.015231,167 3.135494,169 3.135494,171 3.135494,172 3.135494,173 1.098612,176 2.154665,178 2.847812,182 2.442347,188 2.154665,189 1.931521,196 3.135494,199 2.442347,204 3.135494,205 3.135494,206 2.847812,207 3.540959,210 2.847812,211 2.442347,212 2.847812,217 2.442347,221 2.624669,223 0.342286,228 1.931521,236 2.442347,237 3.135494,238 2.847812,252 2.288196,254 3.135494,264 2.624669,265 2.624669,268 3.135494,270 3.540959,271 3.540959,273 2.624669,278 1.7492,282 0.342286,285 2.847812,293 2.036882,294 2.288196,295 2.442347,302 1.056053,303 3.135494,305 1.931521,306 3.135494,318 2.847812,319 1.931521,324 1.289668,326 1.595049,329 1.836211,337 3.135494,383 0.800119,388 1.931521,394 3.135494,395 3.135494,399 2.288196,406 2.847812,408 0.245122,410 1.595049,412 3.540959,415 1.836211,427 2.154665,428 2.442347,429 2.624669,432 3.135494,436 2.847812,437 1.836211,441 3.540959,443 0.342286,444 0.472906,445 2.847812,449 1.7492,451 2.847812,459 0.342286,460 2.442347,461 0.623189,464 2.847812,465 2.624669,468 3.540959,476 0.768371,477 0.075223,481 3.135494,484 1.931521,485 2.442347,491 2.624669,493 0.342286,497 2.847812,499 3.135494,500 0.93827,502 2.154665,506 3.540959,508 3.540959,509 2.442347,510 2.442347,511 3.540959,513 2.847812,519 2.442347,523 2.154665,529 3.135494,532 1.931521,534 0.97601,537 1.7492,538 2.624669,548 2.288196,555 2.624669,557 2.847812,559 2.847812,560 3.135494,561 0.832909,562 2.847812,563 0.768371,570 2.036882,572 0.362905,578 1.836211,583 2.624669,586 3.135494,589 2.624669,595 2.036882,599 3.540959,604 2.624669,611 0.866811,618 2.036882,623 2.288196,630 2.847812,635 2.624669,640 2.442347,641 2.154665,642 3.135494,650 3.135494,654 2.847812,668 2.624669,682 2.288196,683 2.847812,684 0.322083,688 2.036882,689 2.288196,690 3.540959,691 1.526056,703 2.624669,704 2.442347,705 2.288196,706 2.847812,707 2.624669,711 2.154665,717 2.847812,718 3.540959,723 2.847812,724 0.97601,725 2.624669,726 2.624669,727 3.540959,732 2.624669,736 3.540959,737 2.624669,741 2.624669,743 2.847812,744 2.624669,751 2.847812,755 2.442347,759 1.836211,760 2.847812,761 2.624669,769 1.931521,773 2.624669,774 1.931521,783 2.154665,790 2.624669,793 2.442347,794 2.624669,795 2.847812,805 2.847812,806 0.520534,821 3.135494,829 2.288196,835 1.526056,839 3.540959,845 0.768371,847 2.847812,849 0.93827,850 2.442347,857 0.97601,863 1.931521,870 3.135494,873 2.442347,875 2.154665,878 2.847812,879 3.135494,881 2.442347,883 2.847812,884 2.442347,892 2.442347,908 2.288196,911 2.624669,918 0.342286,921 2.624669,922 1.931521,927 3.135494,929 1.015231,932 3.135494,936 3.135494,938 2.442347,946 2.624669,950

3.135494,952 0.93827,962 2.036882,963 2.288196,965 2.847812,970 2.288196,971 2.847812,972 2.847812,973 2.036882,974 3.540959,978 2.847812,987 2.154665,989 2.624669,991 1.595049,996 0.832909,1002 2.036882,1007 3.135494,1009 2.847812,1011 1.836211,1015 3.540959,1018 0.97601,1019 0.768371,1024 0.570545,1026 1.7492,1029 2.288196,1030 2.154665,1031 2.288196,1032 2.624669,1033 2.442347,1034 1.836211,1042 0.014599,1043 3.135494,1045 2.288196,1051 3.540959,1061 0.707746,1062 0.800119,1063 2.288196,1065 0.901902,1080 2.288196,1083 0.97601,1093 1.836211,1094 2.624669,1096 0.678758,1102 0.123233,1105 2.288196}

Cluster 1: {0 'Nonlinear Sciences httpsarxivorgarchivenlin Page 25',4 1.238374,25 1.189584,28 1.189584,34 1.238374,38 2.847812,42 1.238374,51 1.289668,52 1.931521,53 1.238374,80 1.189584,84 1.189584,92 1.143064,93 2.624669,115 1.238374,119 0.342286,128 2.624669,143 1.189584,159 1.189584,160 3.135494,164 1.238374,166 1.238374,174 2.847812,181 1.189584,184 1.238374,191 1.143064,216 2.847812,218 1.238374,222 2.847812,224 1.238374,226 1.238374,231 2.847812,238 2.847812,279 2.288196,281 0.449917,288 1.189584,319 1.931521,348 1.238374,376 1.289668,384 0.427444,391 1.143064,408 0.245122,415 1.836211,431 1.238374,433 0.263815,466 1.238374,473 1.189584,477 0.075223,478 1.189584,496 1.238374,509 2.442347,521 1.238374,530 1.143064,533 1.238374,534 0.97601,559 2.847812,568 1.143064,569 2.288196,592 1.189584,632 2.624669,643 2.847812,646 1.143064,648 1.143064,649 1.143064,676 1.143064,686 1.189584,697 1.143064,700 1.238374,701 2.847812,708 0.93827,713 1.189584,721 1.189584,722 2.624669,723 2.847812,733 1.098612,737 2.624669,739 1.189584,740 1.143064,757 0.472906,769 1.931521,774 1.931521,781 0.472906,782 1.238374,784 0.342286,796 0.901902,801 2.624669,804 0.362905,833 1.189584,837 0.832909,844 1.289668,849 0.93827,855 1.595049,861 1.098612,874 2.442347,875 2.154665,892 2.442347,903 1.056053,926 0.472906,930 1.098612,934 0.263815,941 1.238374,942 0.383959,947 1.189584,963 2.288196,966 1.143064,968 0.97601,977 1.015231,984 2.288196,985 0.866811,987 2.154665,989 2.624669,1011 1.836211,1018 0.97601,1039 2.154665,1042 0.014599,1044 1.143064,1046 3.135494,1075 1.098612,1086 0.472906,1097 1.238374,1102 0.123233,1103 1.015231,1104 1.143064,1108 0.449917,1111 0.800119}

Cluster 2: {0 'To submit an article e-print repository63',2 0.342286,28 1.189584,72 0.362905,97 3.135494,99 3.135494,101 3.540959,117 0.362905,122 3.135494,137 1.461518,144 0.342286,149 1.931521,170 3.540959,173 1.098612,202 3.135494,215 3.540959,217 2.442347,223 0.342286,232 3.135494,249 3.540959,259 2.624669,282 0.342286,286 0.866811,297 2.624669,308 2.442347,309 3.135494,310 3.540959,314 3.540959,321 1.015231,324 1.289668,328 2.624669,329 1.836211,332 3.540959,333 2.624669,336 2.288196,340 3.540959,345 3.540959,350 2.847812,356 3.540959,358 3.540959,361 2.624669,362 0.901902,380 1.931521,383 0.800119,387 3.540959,400 2.624669,402 3.540959,408 0.245122,410 1.595049,418 3.540959,421 0.97601,422 0.901902,423 2.847812,424 0.768371,433 0.263815,439 2.442347,440 2.442347,443 0.342286,444 0.472906,447 2.847812,449 1.7492,453 2.624669,456 3.540959,457 1.931521,459 0.342286,460 2.442347,461 0.623189,463 2.847812,465 2.624669,467 0.97601,476 0.768371,477 0.075223,479 0.737599,481 3.135494,482 2.847812,483 3.540959,487 3.135494,488 1.056053,489 1.056053,493 0.342286,513 2.847812,531 2.847812,534 0.97601,537 1.7492,539 0.97601,540 1.015231,561 0.832909,562 2.847812,566 2.442347,567 3.540959,570 2.036882,572 0.362905,573 0.866811,591 2.847812,595 2.036882,606 3.135494,609 2.847812,614 3.540959,615 2.847812,616 2.442347,626 2.288196,629 2.624669,633 3.135494,639 2.624669,641 2.154665,645 2.154665,660 2.154665,663 1.238374,682 2.288196,684 0.322083,688 2.036882,691 1.526056,699 2.442347,711 2.154665,712 2.442347,720 2.442347,724 0.97601,728 3.135494,730 2.442347,735 2.847812,738 0.737599,741 2.624669,743 2.847812,745 2.624669,759 1.836211,768 3.540959,779 1.289668,784 0.342286,787 2.442347,789 0.93827,791 3.135494,793 2.442347,796 0.901902,799 2.847812,806 0.520534,807 3.135494,822 2.442347,824 0.901902,825 2.847812,831 2.442347,834 2.624669,835 1.526056,837 0.832909,840 2.154665,857 0.97601,861 1.098612,865 3.135494,872 2.847812,873 2.442347,875 2.154665,881 2.442347,882 2.847812,884 2.442347,885 2.847812,898 2.442347,899 2.847812,906 1.015231,908 2.288196,916 1.015231,917 3.135494,918 0.342286,921 2.624669,928 1.015231,929 1.015231,934 0.263815,935 2.847812,945 2.847812,946 2.624669,948 1.931521,952 0.93827,962 2.036882,968 0.97601,970 2.288196,975 2.624669,976 2.847812,992 0.707746,993 2.288196,995 2.442347,996 0.832909,1002 2.036882,1010 2.154665,1018 0.97601,1020 1.461518,1024 0.570545,1026 1.7492,1027 2.847812,1028 2.847812,1031 2.288196,1032 2.624669,1034 1.836211,1038 2.154665,1039 2.154665,1041 0.97601,1042 0.014599,1053 2.624669,1055 2.624669,1061 0.707746,1062 0.800119,1063 2.288196,1064 2.442347,1067 3.135494,1069 3.135494,1071 2.624669,1072 2.847812,1080 2.288196,1083 0.97601,1093 1.836211,1094 2.624669,1096 0.678758,1101 2.847812,1102 0.123233,1105 2.288196,1111 0.800119}

Missing values globally replaced with mean/mode

Time taken to build model (full training data) : 0.06 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 2 (3%)

1 42 (61%)

2 25 (36%)

K-Means = 5 dla danych z filtrem IDFTransform:

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: dokumenty-weka.filters.unsupervised.attribute.StringToNominal-Rfirst-weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-I-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"-weka.filters.unsupervised.attribute.Remove-R2-5

Instances: 69

Attributes: 1112

[list of attributes omitted]

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 3

Within cluster sum of squared errors: 2560.4531620553344

Initial starting points (random):

Cluster 0: {0 'Computer Science Subject Areas and Moderators e-40',1 0.832909,2 0.342286,3 3.540959,5 3.135494,8 3.540959,18 2.624669,20 2.036882,22 2.847812,27 3.135494,37 2.624669,38 2.847812,45 2.847812,54 3.135494,55 3.135494,57 2.442347,58 3.135494,59 2.624669,64 3.135494,65 2.288196,66 2.442347,67 1.931521,68 2.442347,70 2.288196,72 0.362905,76 0.97601,81 2.154665,83 2.624669,90 2.847812,91 3.135494,94 3.540959,100 3.135494,108 2.036882,113 3.540959,117 0.362905,118 2.154665,119 0.342286,124 3.540959,129 1.526056,130 2.847812,133 3.135494,134 2.847812,142 2.847812,144 0.342286,145 3.135494,146 2.624669,149 1.931521,150 2.624669,152 1.7492,156 3.135494,158 1.015231,167 3.135494,169 3.135494,171 3.135494,172 3.135494,173 1.098612,176 2.154665,178 2.847812,182 2.442347,188 2.154665,189 1.931521,196 3.135494,199 2.442347,204 3.135494,205 3.135494,206 2.847812,207 3.540959,210 2.847812,211 2.442347,212 2.847812,217 2.442347,221 2.624669,223 0.342286,228 1.931521,236 2.442347,237 3.135494,238 2.847812,252 2.288196,254 3.135494,264 2.624669,265 2.624669,268 3.135494,270 3.540959,271 3.540959,273 2.624669,278 1.7492,282 0.342286,285 2.847812,293 2.036882,294 2.288196,295 2.442347,302 1.056053,303 3.135494,305 1.931521,306 3.135494,318 2.847812,319 1.931521,324 1.289668,326 1.595049,329 1.836211,337 3.135494,383 0.800119,388 1.931521,394 3.135494,395 3.135494,399 2.288196,406 2.847812,408 0.245122,410 1.595049,412 3.540959,415 1.836211,427 2.154665,428 2.442347,429 2.624669,432 3.135494,436 2.847812,437 1.836211,441 3.540959,443 0.342286,444 0.472906,445 2.847812,449 1.7492,451 2.847812,459 0.342286,460 2.442347,461 0.623189,464 2.847812,465 2.624669,468 3.540959,476 0.768371,477 0.075223,481 3.135494,484 1.931521,485 2.442347,491 2.624669,493 0.342286,497 2.847812,499 3.135494,500 0.93827,502 2.154665,506 3.540959,508 3.540959,509 2.442347,510 2.442347,511 3.540959,513 2.847812,519 2.442347,523 2.154665,529 3.135494,532 1.931521,534 0.97601,537 1.7492,538 2.624669,548 2.288196,555 2.624669,557 2.847812,559 2.847812,560 3.135494,561 0.832909,562 2.847812,563 0.768371,570 2.036882,572 0.362905,578 1.836211,583 2.624669,586 3.135494,589 2.624669,595 2.036882,599 3.540959,604 2.624669,611 0.866811,618 2.036882,623 2.288196,630 2.847812,635 2.624669,640 2.442347,641 2.154665,642 3.135494,650 3.135494,654 2.847812,668 2.624669,682 2.288196,683 2.847812,684 0.322083,688 2.036882,689 2.288196,690 3.540959,691 1.526056,703 2.624669,704 2.442347,705 2.288196,706 2.847812,707 2.624669,711 2.154665,717 2.847812,718 3.540959,723 2.847812,724 0.97601,725 2.624669,726 2.624669,727 3.540959,732 2.624669,736 3.540959,737 2.624669,741 2.624669,743 2.847812,744 2.624669,751 2.847812,755 2.442347,759 1.836211,760 2.847812,761 2.624669,769 1.931521,773 2.624669,774 1.931521,783 2.154665,790 2.624669,793 2.442347,794 2.624669,795 2.847812,805 2.847812,806 0.520534,821 3.135494,829 2.288196,835 1.526056,839 3.540959,845 0.768371,847 2.847812,849 0.93827,850 2.442347,857 0.97601,863 1.931521,870 3.135494,873 2.442347,875 2.154665,878 2.847812,879 3.135494,881 2.442347,883 2.847812,884 2.442347,892 2.442347,908 2.288196,911 2.624669,918 0.342286,921 2.624669,922 1.931521,927 3.135494,929 1.015231,932 3.135494,936 3.135494,938 2.442347,946 2.624669,950 3.135494,952 0.93827,962 2.036882,963 2.288196,965 2.847812,970 2.288196,971 2.847812,972 2.847812,973 2.036882,974 3.540959,978 2.847812,987 2.154665,989 2.624669,991 1.595049,996 0.832909,1002 2.036882,1007 3.135494,1009 2.847812,1011 1.836211,1015 3.540959,1018 0.97601,1019 0.768371,1024 0.570545,1026 1.7492,1029 2.288196,1030 2.154665,1031 2.288196,1032 2.624669,1033 2.442347,1034 1.836211,1042 0.014599,1043 3.135494,1045 2.288196,1051 3.540959,1061 0.707746,1062 0.800119,1063 2.288196,1065 0.901902,1080 2.288196,1083 0.97601,1093 1.836211,1094 2.624669,1096 0.678758,1102 0.123233,1105 2.288196}

Cluster 1: {0 'Nonlinear Sciences httpsarxivorgarchivenlin Page 25',4 1.238374,25 1.189584,28 1.189584,34 1.238374,38 2.847812,42 1.238374,51 1.289668,52 1.931521,53 1.238374,80 1.189584,84 1.189584,92 1.143064,93 2.624669,115 1.238374,119 0.342286,128 2.624669,143 1.189584,159 1.189584,160 3.135494,164 1.238374,166 1.238374,174 2.847812,181 1.189584,184 1.238374,191 1.143064,216 2.847812,218 1.238374,222 2.847812,224 1.238374,226 1.238374,231 2.847812,238 2.847812,279 2.288196,281 0.449917,288 1.189584,319 1.931521,348 1.238374,376 1.289668,384 0.427444,391 1.143064,408 0.245122,415 1.836211,431 1.238374,433 0.263815,466 1.238374,473 1.189584,477 0.075223,478 1.189584,496 1.238374,509 2.442347,521 1.238374,530 1.143064,533 1.238374,534 0.97601,559 2.847812,568 1.143064,569 2.288196,592 1.189584,632 2.624669,643 2.847812,646 1.143064,648 1.143064,649 1.143064,676 1.143064,686 1.189584,697 1.143064,700 1.238374,701 2.847812,708 0.93827,713 1.189584,721 1.189584,722 2.624669,723 2.847812,733 1.098612,737 2.624669,739 1.189584,740 1.143064,757 0.472906,769 1.931521,774 1.931521,781 0.472906,782 1.238374,784 0.342286,796 0.901902,801 2.624669,804 0.362905,833 1.189584,837 0.832909,844 1.289668,849 0.93827,855 1.595049,861 1.098612,874 2.442347,875 2.154665,892 2.442347,903 1.056053,926 0.472906,930 1.098612,934 0.263815,941 1.238374,942 0.383959,947 1.189584,963 2.288196,966 1.143064,968 0.97601,977 1.015231,984 2.288196,985 0.866811,987 2.154665,989 2.624669,1011 1.836211,1018 0.97601,1039 2.154665,1042 0.014599,1044 1.143064,1046 3.135494,1075 1.098612,1086 0.472906,1097 1.238374,1102 0.123233,1103 1.015231,1104 1.143064,1108 0.449917,1111 0.800119}

Cluster 2: {0 'To submit an article e-print repository63',2 0.342286,28 1.189584,72 0.362905,97 3.135494,99 3.135494,101 3.540959,117 0.362905,122 3.135494,137 1.461518,144 0.342286,149 1.931521,170 3.540959,173 1.098612,202 3.135494,215 3.540959,217 2.442347,223 0.342286,232 3.135494,249 3.540959,259 2.624669,282 0.342286,286 0.866811,297 2.624669,308 2.442347,309 3.135494,310 3.540959,314 3.540959,321 1.015231,324 1.289668,328 2.624669,329 1.836211,332 3.540959,333 2.624669,336 2.288196,340 3.540959,345 3.540959,350 2.847812,356 3.540959,358 3.540959,361 2.624669,362 0.901902,380 1.931521,383 0.800119,387 3.540959,400 2.624669,402 3.540959,408 0.245122,410 1.595049,418 3.540959,421 0.97601,422 0.901902,423 2.847812,424 0.768371,433 0.263815,439 2.442347,440 2.442347,443 0.342286,444 0.472906,447 2.847812,449 1.7492,453 2.624669,456 3.540959,457 1.931521,459 0.342286,460 2.442347,461 0.623189,463 2.847812,465 2.624669,467 0.97601,476 0.768371,477 0.075223,479 0.737599,481 3.135494,482 2.847812,483 3.540959,487 3.135494,488 1.056053,489 1.056053,493 0.342286,513 2.847812,531 2.847812,534 0.97601,537 1.7492,539 0.97601,540 1.015231,561 0.832909,562 2.847812,566 2.442347,567 3.540959,570 2.036882,572 0.362905,573 0.866811,591 2.847812,595 2.036882,606 3.135494,609 2.847812,614 3.540959,615 2.847812,616 2.442347,626 2.288196,629 2.624669,633 3.135494,639 2.624669,641 2.154665,645 2.154665,660 2.154665,663 1.238374,682 2.288196,684 0.322083,688 2.036882,691 1.526056,699 2.442347,711 2.154665,712 2.442347,720 2.442347,724 0.97601,728 3.135494,730 2.442347,735 2.847812,738 0.737599,741 2.624669,743 2.847812,745 2.624669,759 1.836211,768 3.540959,779 1.289668,784 0.342286,787 2.442347,789 0.93827,791 3.135494,793 2.442347,796 0.901902,799 2.847812,806 0.520534,807 3.135494,822 2.442347,824 0.901902,825 2.847812,831 2.442347,834 2.624669,835 1.526056,837 0.832909,840 2.154665,857 0.97601,861 1.098612,865 3.135494,872 2.847812,873 2.442347,875 2.154665,881 2.442347,882 2.847812,884 2.442347,885 2.847812,898 2.442347,899 2.847812,906 1.015231,908 2.288196,916 1.015231,917 3.135494,918 0.342286,921 2.624669,928

1.015231,929 1.015231,934 0.263815,935 2.847812,945 2.847812,946 2.624669,948 1.931521,952 0.93827,962 2.036882,968 0.97601,970 2.288196,975 2.624669,976 2.847812,992 0.707746,993 2.288196,995 2.442347,996 0.832909,1002 2.036882,1010 2.154665,1018 0.97601,1020 1.461518,1024 0.570545,1026 1.7492,1027 2.847812,1028 2.847812,1031 2.288196,1032 2.624669,1034 1.836211,1038 2.154665,1039 2.154665,1041 0.97601,1042 0.014599,1053 2.624669,1055 2.624669,1061 0.707746,1062 0.800119,1063 2.288196,1064 2.442347,1067 3.135494,1069 3.135494,1071 2.624669,1072 2.847812,1080 2.288196,1083 0.97601,1093 1.836211,1094 2.624669,1096 0.678758,1101 2.847812,1102 0.123233,1105 2.288196,1111 0.800119}

Cluster 3: {0 'Contacting e-print repository httpsarxivorghelp68',2 0.342286,72 0.362905,97 3.135494,99 3.135494,114 3.135494,117 0.362905,122 3.135494,137 1.461518,144 0.342286,173 1.098612,194 2.847812,202 3.135494,217 2.442347,223 0.342286,247 2.154665,282 0.342286,287 3.135494,336 2.288196,380 1.931521,399 2.288196,408 0.245122,443 0.342286,444 0.472906,449 1.7492,454 2.624669,459 0.342286,461 0.623189,479 0.737599,482 2.847812,493 0.342286,522 2.036882,546 3.540959,550 2.288196,570 2.036882,572 0.362905,584 2.442347,595 2.036882,617 2.624669,639 2.624669,641 2.154665,663 1.238374,677 3.540959,684 0.322083,691 1.526056,699 2.442347,724 0.97601,759 1.836211,766 2.847812,767 2.847812,776 2.442347,779 1.289668,804 0.362905,835 1.526056,837 0.832909,843 0.866811,857 0.97601,860 2.288196,869 1.015231,894 2.847812,898 2.442347,901 2.442347,911 2.624669,918 0.342286,920 2.847812,924 2.847812,929 1.015231,934 0.263815,973 2.036882,991 1.595049,992 0.707746,1012 2.847812,1020 1.461518,1024 0.570545,1031 2.288196,1038 2.154665,1039 2.154665,1042 0.014599,1061 0.707746,1083 0.97601,1102 0.123233,1111 0.800119}

Cluster 4: {0 ' Scientific Advisory Board e-print repository57',2 0.342286,14 2.154665,19 0.707746,24 2.442347,31 3.540959,32 2.288196,39 1.931521,41 2.154665,43 3.540959,48 3.135494,59 2.624669,60 3.135494,67 1.931521,68 2.442347,72 0.362905,75 3.540959,83 2.624669,89 2.847812,117 0.362905,137 1.461518,144 0.342286,152 1.7492,154 2.624669,173 1.098612,189 1.931521,193 2.624669,213 1.836211,223 0.342286,244 2.624669,245 1.595049,253 2.624669,255 1.931521,260 3.540959,265 2.624669,278 1.7492,279 2.288196,280 2.036882,282 0.342286,294 2.288196,298 2.624669,305 1.931521,324 1.289668,325 2.442347,369 4.234107,434 0.97601,443 0.342286,444 0.472906,459 0.342286,470 2.154665,477 0.075223,486 2.624669,493 0.342286,510 2.442347,528 2.442347,534 0.97601,541 3.540959,572 0.362905,573 0.866811,578 1.836211,618 2.036882,626 2.288196,652 2.847812,663 1.238374,684 0.322083,712 2.442347,724 0.97601,726 2.624669,756 2.154665,765 1.836211,776 2.442347,779 1.289668,818 4.234107,826 3.135494,847 2.847812,855 1.595049,857 0.97601,862 2.624669,883 2.847812,891 2.154665,918 0.342286,923 3.135494,929 1.015231,934 0.263815,938 2.442347,953 2.036882,985 0.866811,1010 2.154665,1020 1.461518,1024 0.570545,1042 0.014599,1083 0.97601,1093 1.836211,1102 0.123233,1106 0.866811}

Missing values globally replaced with mean/mode

Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 2 (3%)

1 20 (29%)

2 2 (3%)

3 23 (33%)

4 22 (32%)

K-Means = 3 dla danych z filtrem TFTransform:

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: dokumenty-weka.filters.unsupervised.attribute.StringToNominal-Rfirst-weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-T-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"-weka.filters.unsupervised.attribute.Remove-R2-5-weka.filters.unsupervised.attribute.StringToWordVector-Rfirst-last-W1000-prune-rate-1.0-T-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"

Instances: 69

Attributes: 1112

[list of attributes omitted]

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 6

Within cluster sum of squared errors: 5275.052380952386

Initial starting points (random):

Cluster 0: {0 'Computer Science Subject Areas and Moderators e-40',1 0.693147,2 0.693147,3 0.693147,5 0.693147,8 0.693147,10 0.693147,11 0.693147,13 0.693147,18 0.693147,20 0.693147,22 0.693147,27 0.693147,36 0.693147,37 0.693147,38 0.693147,40 0.693147,45 0.693147,54 0.693147,55 0.693147,57 0.693147,58 0.693147,59 0.693147,61 0.693147,64 0.693147,65 0.693147,66 0.693147,67 0.693147,68 0.693147,70 0.693147,71 0.693147,72 0.693147,76 0.693147,79 0.693147,81 0.693147,83 0.693147,90 0.693147,91 0.693147,94 0.693147,100 0.693147,108 0.693147,113 0.693147,117 0.693147,118 0.693147,119 0.693147,120 0.693147,123 0.693147,124 0.693147,129 0.693147,130 0.693147,133 0.693147,134 0.693147,136 0.693147,142 0.693147,144 0.693147,145 0.693147,146 0.693147,147 0.693147,149 0.693147,150 0.693147,152 0.693147,156 0.693147,158 0.693147,163 0.693147,167 0.693147,169 0.693147,171 0.693147,172 0.693147,173 0.693147,175 0.693147,176 0.693147,178 0.693147,182 0.693147,188 0.693147,189 0.693147,196 0.693147,199 0.693147,204 0.693147,205 0.693147,206 0.693147,207 0.693147,210 0.693147,211 0.693147,212 0.693147,217 0.693147,221 0.693147,223 0.693147,228 0.693147,235 0.693147,236 0.693147,237 0.693147,238 0.693147,248 0.693147,250 0.693147,252 0.693147,254 0.693147,263 0.693147,264 0.693147,265 0.693147,268 0.693147,270 0.693147,271 0.693147,273 0.693147,278 0.693147,282 0.693147,285 0.693147,292 0.693147,293 0.693147,294 0.693147,295 0.693147,302 0.693147,303 0.693147,305 0.693147,306 0.693147,312 0.693147,318 0.693147,319 0.693147,322 0.693147,324 0.693147,326 0.693147,329 0.693147,335 0.693147,337 0.693147,341 0.693147,346 0.693147,354 0.693147,367 0.693147,368 0.693147,379 0.693147,383 0.693147,388 0.693147,390 0.693147,392 0.693147,394 0.693147,395 0.693147,399 0.693147,406 0.693147,408 0.693147,410 0.693147,411 0.693147,412 0.693147,414 0.693147,415 0.693147,419 0.693147,427 0.693147,428 0.693147,429 0.693147,430 0.693147,432 0.693147,435 0.693147,436 0.693147,437 0.693147,441 0.693147,442 0.693147,443 0.693147,444 0.693147,445 0.693147,446 0.693147,449 0.693147,451 0.693147,452 0.693147,459 0.693147,460 0.693147,461 0.693147,464 0.693147,465 0.693147,468 0.693147,476 0.693147,477 0.693147,481 0.693147,484 0.693147,485 0.693147,491 0.693147,492 0.693147,493 0.693147,497 0.693147,499 0.693147,500 0.693147,502 0.693147,506 0.693147,508 0.693147,509 0.693147,510 0.693147,511 0.693147,513 0.693147,515 0.693147,519 0.693147,523 0.693147,525 0.693147,526 0.693147,529 0.693147,532 0.693147,534 0.693147,537 0.693147,538 0.693147,548 0.693147,555 0.693147,556 0.693147,557 0.693147,559 0.693147,560 0.693147,561 0.693147,562 0.693147,563 0.693147,570 0.693147,572 0.693147,578 0.693147,579 0.693147,583 0.693147,586 0.693147,589 0.693147,595 0.693147,599 0.693147,604 0.693147,611 0.693147,612 0.693147,618 0.693147,623 0.693147,627 0.693147,630 0.693147,631 0.693147,634 0.693147,635 0.693147,640 0.693147,641 0.693147,642 0.693147,650 0.693147,654 0.693147,655 0.693147,661 0.693147,662 0.693147,668 0.693147,674 0.693147,678 0.693147,681 0.693147,682 0.693147,683 0.693147,684 0.693147,687 0.693147,688 0.693147,689 0.693147,690 0.693147,691 0.693147,695 0.693147,698 0.693147,703 0.693147,704 0.693147,705 0.693147,706 0.693147,707 0.693147,710 0.693147,711 0.693147,717 0.693147,718 0.693147,723 0.693147,724 0.693147,725 0.693147,726 0.693147,727 0.693147,732 0.693147,736 0.693147,737 0.693147,741 0.693147,743 0.693147,744 0.693147,751 0.693147,752 0.693147,755 0.693147,759 0.693147,760 0.693147,761 0.693147,764 0.693147,769 0.693147,773 0.693147,774 0.693147,783 0.693147,790 0.693147,792 0.693147,793 0.693147,794 0.693147,795 0.693147,805 0.693147,806 0.693147,812 0.693147,816 0.693147,820 0.693147,821 0.693147,829 0.693147,830 0.693147,832 0.693147,835 0.693147,839 0.693147,841 0.693147,845 0.693147,847 0.693147,849 0.693147,850 0.693147,853 0.693147,857 0.693147,863 0.693147,870 0.693147,873 0.693147,875 0.693147,878 0.693147,879 0.693147,881 0.693147,883 0.693147,884 0.693147,892 0.693147,900 0.693147,904 0.693147,907 0.693147,908 0.693147,911 0.693147,918 0.693147,921 0.693147,922 0.693147,927 0.693147,929 0.693147,932 0.693147,936 0.693147,938 0.693147,946 0.693147,950 0.693147,952 0.693147,962 0.693147,963 0.693147,965 0.693147,970 0.693147,971 0.693147,972 0.693147,973 0.693147,974 0.693147,978 0.693147,983 0.693147,987 0.693147,989 0.693147,991 0.693147,996 0.693147,1002 0.693147,1004 0.693147,1007 0.693147,1008 0.693147,1009 0.693147,1011 0.693147,1015 0.693147,1018 0.693147,1019 0.693147,1024 0.693147,1025 0.693147,1026 0.693147,1029 0.693147,1030 0.693147,1031 0.693147,1032 0.693147,1033 0.693147,1034 0.693147,1035 0.693147,1042 0.693147,1043 0.693147,1045 0.693147,1048 0.693147,1050 0.693147,1051 0.693147,1059 0.693147,1061 0.693147,1062 0.693147,1063 0.693147,1065 0.693147,1080 0.693147,1082 0.693147,1083 0.693147,1093 0.693147,1094 0.693147,1096 0.693147,1102 0.693147,1105 0.693147,1110 0.693147}

Cluster 1: {0 'Nonlinear Sciences httpsarxivorgarchivenlin Page 25',4 0.693147,10 0.693147,11 0.693147,13 0.693147,25 0.693147,28 0.693147,34 0.693147,36 0.693147,38 0.693147,40 0.693147,42 0.693147,51 0.693147,52 0.693147,53 0.693147,61 0.693147,71 0.693147,79 0.693147,80 0.693147,84 0.693147,92 0.693147,93 0.693147,115 0.693147,119 0.693147,120 0.693147,123 0.693147,128 0.693147,136 0.693147,143 0.693147,147 0.693147,159 0.693147,160 0.693147,163 0.693147,164 0.693147,166 0.693147,174 0.693147,175 0.693147,181 0.693147,184 0.693147,191 0.693147,216 0.693147,218 0.693147,222 0.693147,224 0.693147,226 0.693147,231 0.693147,235 0.693147,238 0.693147,248 0.693147,250 0.693147,263 0.693147,279 0.693147,281 0.693147,288 0.693147,292 0.693147,312 0.693147,319 0.693147,322 0.693147,335 0.693147,341 0.693147,346 0.693147,348 0.693147,354 0.693147,367 0.693147,368 0.693147,376 0.693147,379 0.693147,384 0.693147,390 0.693147,391 0.693147,392 0.693147,408 0.693147,411 0.693147,414 0.693147,415 0.693147,419 0.693147,430 0.693147,431 0.693147,433 0.693147,435 0.693147,442 0.693147,446 0.693147,452 0.693147,466 0.693147,473 0.693147,477 0.693147,478 0.693147,492 0.693147,496 0.693147,509 0.693147,515 0.693147,521 0.693147,525 0.693147,526 0.693147,530 0.693147,533 0.693147,534 0.693147,556 0.693147,559 0.693147,568 0.693147,569 0.693147,579 0.693147,592 0.693147,612 0.693147,627 0.693147,631 0.693147,632 0.693147,634 0.693147,643 0.693147,646 0.693147,648 0.693147,649 0.693147,655 0.693147,661 0.693147,662 0.693147,674 0.693147,676 0.693147,678 0.693147,681 0.693147,686 0.693147,687 0.693147,695 0.693147,697 0.693147,698 0.693147,700 0.693147,701 0.693147,708 0.693147,710 0.693147,713 0.693147,721 0.693147,722 0.693147,723 0.693147,733 0.693147,737 0.693147,739 0.693147,740 0.693147,752 0.693147,757 0.693147,764 0.693147,769 0.693147,774 0.693147,781 0.693147,782 0.693147,784 0.693147,792 0.693147,796 0.693147,801 0.693147,804 0.693147,812 0.693147,816 0.693147,820 0.693147,830 0.693147,832 0.693147,833 0.693147,837 0.693147,841 0.693147,844 0.693147,849 0.693147,853 0.693147,855 0.693147,861 0.693147,874 0.693147,875 0.693147,892 0.693147,900 0.693147,903 0.693147,904 0.693147,907 0.693147,926 0.693147,930 0.693147,934 0.693147,941 0.693147,942 0.693147,947 0.693147,963 0.693147,966 0.693147,968 0.693147,977 0.693147,983 0.693147,984 0.693147,985 0.693147,987 0.693147,989 0.693147,1004 0.693147,1008 0.693147,1011 0.693147,1018 0.693147,1025 0.693147,1035 0.693147,1039 0.693147,1042 0.693147,1044 0.693147,1046 0.693147,1048 0.693147,1050 0.693147,1059 0.693147,1075 0.693147,1082 0.693147,1086 0.693147,1097 0.693147,1102 0.693147,1103 0.693147,1104 0.693147,1108 0.693147,1110 0.693147,1111 0.693147}

Cluster 2: {0 'To submit an article e-print repository63',2 0.693147,10 0.693147,11 0.693147,13 0.693147,28 0.693147,36 0.693147,40 0.693147,61 0.693147,71 0.693147,72 0.693147,79 0.693147,97 0.693147,99 0.693147,101 0.693147,117 0.693147,120 0.693147,122 0.693147,123 0.693147,136 0.693147,137 0.693147,144 0.693147,147 0.693147,149 0.693147,163 0.693147,170 0.693147,173 0.693147,175 0.693147,202 0.693147,215 0.693147,217 0.693147,223 0.693147,232 0.693147,235 0.693147,248 0.693147,249 0.693147,250 0.693147,259 0.693147,263 0.693147,282 0.693147,286 0.693147,292 0.693147,297 0.693147,308 0.693147,309 0.693147,310 0.693147,312 0.693147,314 0.693147,321 0.693147,322 0.693147,324 0.693147,328 0.693147,329 0.693147,332 0.693147,333 0.693147,335 0.693147,336 0.693147,340 0.693147,341 0.693147,345 0.693147,346 0.693147,350 0.693147,354 0.693147,356 0.693147,358 0.693147,361 0.693147,362 0.693147,367 0.693147,368 0.693147,379 0.693147,380 0.693147,383 0.693147,387 0.693147,390 0.693147,392 0.693147,400 0.693147,402 0.693147,408 0.693147,410 0.693147,411 0.693147,414 0.693147,418 0.693147,419 0.693147,421 0.693147,422 0.693147,423

0.693147,424 0.693147,430 0.693147,433 0.693147,435 0.693147,439 0.693147,440 0.693147,442 0.693147,443 0.693147,444 0.693147,446 0.693147,447
0.693147,449 0.693147,452 0.693147,453 0.693147,456 0.693147,457 0.693147,459 0.693147,460 0.693147,461 0.693147,463 0.693147,465 0.693147,467
0.693147,476 0.693147,477 0.693147,479 0.693147,481 0.693147,482 0.693147,483 0.693147,487 0.693147,488 0.693147,489 0.693147,492 0.693147,493
0.693147,513 0.693147,515 0.693147,525 0.693147,526 0.693147,531 0.693147,534 0.693147,537 0.693147,539 0.693147,540 0.693147,556 0.693147,561
0.693147,562 0.693147,566 0.693147,567 0.693147,570 0.693147,572 0.693147,573 0.693147,579 0.693147,591 0.693147,595 0.693147,606 0.693147,609
0.693147,612 0.693147,614 0.693147,615 0.693147,616 0.693147,626 0.693147,627 0.693147,629 0.693147,631 0.693147,633 0.693147,634 0.693147,639
0.693147,641 0.693147,645 0.693147,655 0.693147,660 0.693147,661 0.693147,662 0.693147,663 0.693147,674 0.693147,678 0.693147,681 0.693147,682
0.693147,684 0.693147,687 0.693147,688 0.693147,691 0.693147,695 0.693147,698 0.693147,699 0.693147,710 0.693147,711 0.693147,712 0.693147,720
0.693147,724 0.693147,728 0.693147,730 0.693147,735 0.693147,738 0.693147,741 0.693147,743 0.693147,745 0.693147,752 0.693147,759 0.693147,764
0.693147,768 0.693147,779 0.693147,784 0.693147,787 0.693147,789 0.693147,791 0.693147,792 0.693147,793 0.693147,796 0.693147,799 0.693147,806
0.693147,807 0.693147,812 0.693147,816 0.693147,820 0.693147,822 0.693147,824 0.693147,825 0.693147,830 0.693147,831 0.693147,832 0.693147,834
0.693147,835 0.693147,837 0.693147,840 0.693147,841 0.693147,853 0.693147,857 0.693147,861 0.693147,865 0.693147,872 0.693147,873 0.693147,875
0.693147,881 0.693147,882 0.693147,884 0.693147,885 0.693147,898 0.693147,899 0.693147,900 0.693147,904 0.693147,906 0.693147,907 0.693147,908
0.693147,916 0.693147,917 0.693147,918 0.693147,921 0.693147,928 0.693147,929 0.693147,934 0.693147,935 0.693147,945 0.693147,946 0.693147,948
0.693147,952 0.693147,962 0.693147,968 0.693147,970 0.693147,975 0.693147,976 0.693147,983 0.693147,992 0.693147,993 0.693147,995 0.693147,996
0.693147,1002 0.693147,1004 0.693147,1008 0.693147,1010 0.693147,1018 0.693147,1020 0.693147,1024 0.693147,1025 0.693147,1026 0.693147,1027
0.693147,1028 0.693147,1031 0.693147,1032 0.693147,1034 0.693147,1035 0.693147,1038 0.693147,1039 0.693147,1041 0.693147,1042 0.693147,1048
0.693147,1050 0.693147,1053 0.693147,1055 0.693147,1059 0.693147,1061 0.693147,1062 0.693147,1063 0.693147,1064 0.693147,1067 0.693147,1069
0.693147,1071 0.693147,1072 0.693147,1080 0.693147,1082 0.693147,1083 0.693147,1093 0.693147,1094 0.693147,1096 0.693147,1101 0.693147,1102
0.693147,1105 0.693147,1110 0.693147,1111 0.693147}

Missing values globally replaced with mean/mode

Time taken to build model (full training data) : 0.07 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 2 (3%)

1 42 (61%)

2 25 (36%)

K-Means = 5 dla danych z filtrem TFTransform:

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: dokumenty-weka.filters.unsupervised.attribute.StringToNominal-Rfirst-weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-T-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"-weka.filters.unsupervised.attribute.Remove-R2-5-weka.filters.unsupervised.attribute.StringToWordVector-Rfirst-last-W1000-prune-rate-1.0-T-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"

Instances: 69

Attributes: 1112

[list of attributes omitted]

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 3

Within cluster sum of squared errors: 2560.4531620553344

Initial starting points (random):

Cluster 0: {0 'Computer Science Subject Areas and Moderators e-40',1 0.693147,2 0.693147,3 0.693147,5 0.693147,8 0.693147,10 0.693147,11 0.693147,13 0.693147,18 0.693147,20 0.693147,22 0.693147,27 0.693147,36 0.693147,37 0.693147,38 0.693147,40 0.693147,45 0.693147,54 0.693147,55 0.693147,57

0.693147,58 0.693147,59 0.693147,61 0.693147,64 0.693147,65 0.693147,66 0.693147,67 0.693147,68 0.693147,70 0.693147,71 0.693147,72 0.693147,76
0.693147,79 0.693147,81 0.693147,83 0.693147,90 0.693147,91 0.693147,94 0.693147,100 0.693147,108 0.693147,113 0.693147,117 0.693147,118
0.693147,119 0.693147,120 0.693147,123 0.693147,124 0.693147,129 0.693147,130 0.693147,133 0.693147,134 0.693147,136 0.693147,142 0.693147,144
0.693147,145 0.693147,146 0.693147,147 0.693147,149 0.693147,150 0.693147,152 0.693147,156 0.693147,158 0.693147,163 0.693147,167 0.693147,169
0.693147,171 0.693147,172 0.693147,173 0.693147,175 0.693147,176 0.693147,178 0.693147,182 0.693147,188 0.693147,189 0.693147,196 0.693147,199
0.693147,204 0.693147,205 0.693147,206 0.693147,207 0.693147,210 0.693147,211 0.693147,212 0.693147,217 0.693147,221 0.693147,223 0.693147,228
0.693147,235 0.693147,236 0.693147,237 0.693147,238 0.693147,248 0.693147,250 0.693147,252 0.693147,254 0.693147,263 0.693147,264 0.693147,265
0.693147,268 0.693147,270 0.693147,271 0.693147,273 0.693147,278 0.693147,282 0.693147,285 0.693147,292 0.693147,293 0.693147,294 0.693147,295
0.693147,302 0.693147,303 0.693147,305 0.693147,306 0.693147,312 0.693147,318 0.693147,319 0.693147,322 0.693147,324 0.693147,326 0.693147,329
0.693147,335 0.693147,337 0.693147,341 0.693147,346 0.693147,354 0.693147,367 0.693147,368 0.693147,379 0.693147,383 0.693147,388 0.693147,390
0.693147,392 0.693147,394 0.693147,395 0.693147,399 0.693147,406 0.693147,408 0.693147,410 0.693147,411 0.693147,412 0.693147,414 0.693147,415
0.693147,419 0.693147,427 0.693147,428 0.693147,429 0.693147,430 0.693147,432 0.693147,435 0.693147,436 0.693147,437 0.693147,441 0.693147,442
0.693147,443 0.693147,444 0.693147,445 0.693147,446 0.693147,449 0.693147,451 0.693147,452 0.693147,459 0.693147,460 0.693147,461 0.693147,464
0.693147,465 0.693147,468 0.693147,476 0.693147,477 0.693147,481 0.693147,484 0.693147,485 0.693147,491 0.693147,492 0.693147,493 0.693147,497
0.693147,499 0.693147,500 0.693147,502 0.693147,506 0.693147,508 0.693147,509 0.693147,510 0.693147,511 0.693147,513 0.693147,515 0.693147,519
0.693147,523 0.693147,525 0.693147,526 0.693147,529 0.693147,532 0.693147,534 0.693147,537 0.693147,538 0.693147,548 0.693147,555 0.693147,556
0.693147,557 0.693147,559 0.693147,560 0.693147,561 0.693147,562 0.693147,563 0.693147,570 0.693147,572 0.693147,578 0.693147,579 0.693147,583
0.693147,586 0.693147,589 0.693147,595 0.693147,599 0.693147,604 0.693147,611 0.693147,612 0.693147,618 0.693147,623 0.693147,627 0.693147,630
0.693147,631 0.693147,634 0.693147,635 0.693147,640 0.693147,641 0.693147,642 0.693147,650 0.693147,654 0.693147,655 0.693147,661 0.693147,662
0.693147,668 0.693147,674 0.693147,678 0.693147,681 0.693147,682 0.693147,683 0.693147,684 0.693147,687 0.693147,688 0.693147,689 0.693147,690
0.693147,691 0.693147,695 0.693147,698 0.693147,703 0.693147,704 0.693147,705 0.693147,706 0.693147,707 0.693147,710 0.693147,711 0.693147,717
0.693147,718 0.693147,723 0.693147,724 0.693147,725 0.693147,726 0.693147,727 0.693147,732 0.693147,736 0.693147,737 0.693147,741 0.693147,743
0.693147,744 0.693147,751 0.693147,752 0.693147,755 0.693147,759 0.693147,760 0.693147,761 0.693147,764 0.693147,769 0.693147,773 0.693147,774
0.693147,783 0.693147,790 0.693147,792 0.693147,793 0.693147,794 0.693147,795 0.693147,805 0.693147,806 0.693147,812 0.693147,816 0.693147,820
0.693147,821 0.693147,829 0.693147,830 0.693147,832 0.693147,835 0.693147,839 0.693147,841 0.693147,845 0.693147,847 0.693147,849 0.693147,850
0.693147,853 0.693147,857 0.693147,863 0.693147,870 0.693147,873 0.693147,875 0.693147,878 0.693147,879 0.693147,881 0.693147,883 0.693147,884
0.693147,892 0.693147,900 0.693147,904 0.693147,907 0.693147,908 0.693147,911 0.693147,918 0.693147,921 0.693147,922 0.693147,927 0.693147,929
0.693147,932 0.693147,936 0.693147,938 0.693147,946 0.693147,950 0.693147,952 0.693147,962 0.693147,963 0.693147,965 0.693147,970 0.693147,971
0.693147,972 0.693147,973 0.693147,974 0.693147,978 0.693147,983 0.693147,987 0.693147,989 0.693147,991 0.693147,996 0.693147,1002 0.693147,1004
0.693147,1007 0.693147,1008 0.693147,1009 0.693147,1011 0.693147,1015 0.693147,1018 0.693147,1019 0.693147,1024 0.693147,1025 0.693147,1026
0.693147,1029 0.693147,1030 0.693147,1031 0.693147,1032 0.693147,1033 0.693147,1034 0.693147,1035 0.693147,1042 0.693147,1043 0.693147,1045
0.693147,1048 0.693147,1050 0.693147,1051 0.693147,1059 0.693147,1061 0.693147,1062 0.693147,1063 0.693147,1065 0.693147,1080 0.693147,1082
0.693147,1083 0.693147,1093 0.693147,1094 0.693147,1096 0.693147,1102 0.693147,1105 0.693147,1110 0.693147}

Cluster 1: {0 'Nonlinear Sciences httpsarxivorgarchivenlin Page 25',4 0.693147,10 0.693147,11 0.693147,13 0.693147,25 0.693147,28 0.693147,34 0.693147,36
0.693147,38 0.693147,40 0.693147,42 0.693147,51 0.693147,52 0.693147,53 0.693147,61 0.693147,71 0.693147,79 0.693147,80 0.693147,84 0.693147,92
0.693147,93 0.693147,115 0.693147,119 0.693147,120 0.693147,123 0.693147,128 0.693147,136 0.693147,143 0.693147,147 0.693147,159 0.693147,160
0.693147,163 0.693147,164 0.693147,166 0.693147,174 0.693147,175 0.693147,181 0.693147,184 0.693147,191 0.693147,216 0.693147,218 0.693147,222
0.693147,224 0.693147,226 0.693147,231 0.693147,235 0.693147,238 0.693147,248 0.693147,250 0.693147,263 0.693147,279 0.693147,281 0.693147,288
0.693147,292 0.693147,312 0.693147,319 0.693147,322 0.693147,335 0.693147,341 0.693147,346 0.693147,348 0.693147,354 0.693147,367 0.693147,368
0.693147,376 0.693147,379 0.693147,384 0.693147,390 0.693147,391 0.693147,392 0.693147,408 0.693147,411 0.693147,414 0.693147,415 0.693147,419
0.693147,430 0.693147,431 0.693147,433 0.693147,435 0.693147,442 0.693147,446 0.693147,452 0.693147,466 0.693147,473 0.693147,477 0.693147,478
0.693147,492 0.693147,496 0.693147,509 0.693147,515 0.693147,521 0.693147,525 0.693147,526 0.693147,530 0.693147,533 0.693147,534 0.693147,556
0.693147,559 0.693147,568 0.693147,569 0.693147,579 0.693147,592 0.693147,612 0.693147,627 0.693147,631 0.693147,632 0.693147,634 0.693147,643
0.693147,646 0.693147,648 0.693147,649 0.693147,655 0.693147,661 0.693147,662 0.693147,674 0.693147,676 0.693147,678 0.693147,681 0.693147,686
0.693147,687 0.693147,695 0.693147,697 0.693147,698 0.693147,700 0.693147,701 0.693147,708 0.693147,710 0.693147,713 0.693147,721 0.693147,722
0.693147,723 0.693147,733 0.693147,737 0.693147,739 0.693147,740 0.693147,752 0.693147,757 0.693147,764 0.693147,769 0.693147,774 0.693147,781
0.693147,782 0.693147,784 0.693147,792 0.693147,796 0.693147,801 0.693147,804 0.693147,812 0.693147,816 0.693147,820 0.693147,830 0.693147,832
0.693147,833 0.693147,837 0.693147,841 0.693147,844 0.693147,849 0.693147,853 0.693147,855 0.693147,861 0.693147,874 0.693147,875 0.693147,892
0.693147,900 0.693147,903 0.693147,904 0.693147,907 0.693147,926 0.693147,930 0.693147,934 0.693147,941 0.693147,942 0.693147,947 0.693147,963
0.693147,966 0.693147,968 0.693147,977 0.693147,983 0.693147,984 0.693147,985 0.693147,987 0.693147,989 0.693147,1004 0.693147,1008
0.693147,1011 0.693147,1018 0.693147,1025 0.693147,1035 0.693147,1039 0.693147,1042 0.693147,1044 0.693147,1046 0.693147,1048 0.693147,1050
0.693147,1059 0.693147,1075 0.693147,1082 0.693147,1086 0.693147,1097 0.693147,1102 0.693147,1103 0.693147,1104 0.693147,1108 0.693147,1110
0.693147,1111 0.693147}

Cluster 2: {0 'To submit an article e-print repository63',2 0.693147,10 0.693147,11 0.693147,13 0.693147,28 0.693147,36 0.693147,40 0.693147,61
0.693147,71 0.693147,72 0.693147,79 0.693147,97 0.693147,99 0.693147,101 0.693147,117 0.693147,120 0.693147,122 0.693147,123 0.693147,136
0.693147,137 0.693147,144 0.693147,147 0.693147,149 0.693147,163 0.693147,170 0.693147,173 0.693147,175 0.693147,202 0.693147,215 0.693147,217
0.693147,223 0.693147,232 0.693147,235 0.693147,248 0.693147,249 0.693147,250 0.693147,259 0.693147,263 0.693147,282 0.693147,286 0.693147,292
0.693147,297 0.693147,308 0.693147,309 0.693147,310 0.693147,312 0.693147,314 0.693147,321 0.693147,322 0.693147,324 0.693147,328 0.693147,329
0.693147,332 0.693147,333 0.693147,335 0.693147,336 0.693147,340 0.693147,341 0.693147,345 0.693147,346 0.693147,350 0.693147,354 0.693147,356
0.693147,358 0.693147,361 0.693147,362 0.693147,367 0.693147,368 0.693147,379 0.693147,380 0.693147,383 0.693147,387 0.693147,390 0.693147,392
0.693147,400 0.693147,402 0.693147,408 0.693147,410 0.693147,411 0.693147,414 0.693147,418 0.693147,419 0.693147,421 0.693147,422 0.693147,423
0.693147,424 0.693147,430 0.693147,433 0.693147,435 0.693147,439 0.693147,440 0.693147,442 0.693147,443 0.693147,444 0.693147,446 0.693147,447
0.693147,449 0.693147,452 0.693147,453 0.693147,456 0.693147,457 0.693147,459 0.693147,460 0.693147,461 0.693147,463 0.693147,465 0.693147,467
0.693147,476 0.693147,477 0.693147,479 0.693147,481 0.693147,482 0.693147,483 0.693147,487 0.693147,488 0.693147,489 0.693147,492 0.693147,493
0.693147,513 0.693147,515 0.693147,525 0.693147,526 0.693147,531 0.693147,534 0.693147,537 0.693147,539 0.693147,540 0.693147,556 0.693147,561
0.693147,562 0.693147,566 0.693147,567 0.693147,570 0.693147,572 0.693147,573 0.693147,579 0.693147,591 0.693147,595 0.693147,606 0.693147,609
0.693147,612 0.693147,614 0.693147,615 0.693147,616 0.693147,626 0.693147,627 0.693147,629 0.693147,631 0.693147,633 0.693147,634 0.693147,639
0.693147,641 0.693147,645 0.693147,655 0.693147,660 0.693147,661 0.693147,662 0.693147,663 0.693147,674 0.693147,678 0.693147,681 0.693147,682
0.693147,684 0.693147,687 0.693147,688 0.693147,691 0.693147,695 0.693147,698 0.693147,699 0.693147,710 0.693147,711 0.693147,712 0.693147,720

0.693147,724 0.693147,728 0.693147,730 0.693147,735 0.693147,738 0.693147,741 0.693147,743 0.693147,745 0.693147,752 0.693147,759 0.693147,764
0.693147,768 0.693147,779 0.693147,784 0.693147,787 0.693147,789 0.693147,791 0.693147,792 0.693147,793 0.693147,796 0.693147,799 0.693147,806
0.693147,807 0.693147,812 0.693147,816 0.693147,820 0.693147,822 0.693147,824 0.693147,825 0.693147,830 0.693147,831 0.693147,832 0.693147,834
0.693147,835 0.693147,837 0.693147,840 0.693147,841 0.693147,853 0.693147,857 0.693147,861 0.693147,865 0.693147,872 0.693147,873 0.693147,875
0.693147,881 0.693147,882 0.693147,884 0.693147,885 0.693147,898 0.693147,899 0.693147,900 0.693147,904 0.693147,906 0.693147,907 0.693147,908
0.693147,916 0.693147,917 0.693147,918 0.693147,921 0.693147,928 0.693147,929 0.693147,934 0.693147,935 0.693147,945 0.693147,946 0.693147,948
0.693147,952 0.693147,962 0.693147,968 0.693147,970 0.693147,975 0.693147,976 0.693147,983 0.693147,992 0.693147,993 0.693147,995 0.693147,996
0.693147,1002 0.693147,1004 0.693147,1008 0.693147,1010 0.693147,1018 0.693147,1020 0.693147,1024 0.693147,1025 0.693147,1026 0.693147,1027
0.693147,1028 0.693147,1031 0.693147,1032 0.693147,1034 0.693147,1035 0.693147,1038 0.693147,1039 0.693147,1041 0.693147,1042 0.693147,1048
0.693147,1050 0.693147,1053 0.693147,1055 0.693147,1059 0.693147,1061 0.693147,1062 0.693147,1063 0.693147,1064 0.693147,1067 0.693147,1069
0.693147,1071 0.693147,1072 0.693147,1080 0.693147,1082 0.693147,1083 0.693147,1093 0.693147,1094 0.693147,1096 0.693147,1101 0.693147,1102
0.693147,1105 0.693147,1110 0.693147,1111 0.693147}

Cluster 3: {0 'Contacting e-print repository httpsarxivorghelp68',2 0.693147,10 0.693147,11 0.693147,13 0.693147,36 0.693147,40 0.693147,61 0.693147,71
0.693147,72 0.693147,79 0.693147,97 0.693147,99 0.693147,114 0.693147,117 0.693147,120 0.693147,122 0.693147,123 0.693147,136 0.693147,137
0.693147,144 0.693147,147 0.693147,163 0.693147,173 0.693147,175 0.693147,194 0.693147,202 0.693147,217 0.693147,223 0.693147,235 0.693147,247
0.693147,248 0.693147,250 0.693147,263 0.693147,282 0.693147,287 0.693147,292 0.693147,312 0.693147,322 0.693147,335 0.693147,336 0.693147,341
0.693147,346 0.693147,354 0.693147,367 0.693147,368 0.693147,379 0.693147,380 0.693147,390 0.693147,392 0.693147,399 0.693147,408 0.693147,411
0.693147,414 0.693147,419 0.693147,430 0.693147,435 0.693147,442 0.693147,443 0.693147,444 0.693147,446 0.693147,449 0.693147,452 0.693147,454
0.693147,459 0.693147,461 0.693147,479 0.693147,482 0.693147,492 0.693147,493 0.693147,515 0.693147,522 0.693147,525 0.693147,526 0.693147,546
0.693147,550 0.693147,556 0.693147,570 0.693147,572 0.693147,579 0.693147,584 0.693147,595 0.693147,612 0.693147,617 0.693147,627 0.693147,631
0.693147,634 0.693147,639 0.693147,641 0.693147,655 0.693147,661 0.693147,662 0.693147,663 0.693147,674 0.693147,677 0.693147,678 0.693147,681
0.693147,684 0.693147,687 0.693147,691 0.693147,695 0.693147,698 0.693147,699 0.693147,710 0.693147,724 0.693147,752 0.693147,759 0.693147,764
0.693147,766 0.693147,767 0.693147,776 0.693147,779 0.693147,792 0.693147,804 0.693147,812 0.693147,816 0.693147,820 0.693147,830 0.693147,832
0.693147,835 0.693147,837 0.693147,841 0.693147,843 0.693147,853 0.693147,857 0.693147,860 0.693147,869 0.693147,894 0.693147,898 0.693147,900
0.693147,901 0.693147,904 0.693147,907 0.693147,911 0.693147,918 0.693147,920 0.693147,924 0.693147,929 0.693147,934 0.693147,973 0.693147,983
0.693147,991 0.693147,992 0.693147,1004 0.693147,1008 0.693147,1012 0.693147,1020 0.693147,1024 0.693147,1025 0.693147,1031 0.693147,1035
0.693147,1038 0.693147,1039 0.693147,1042 0.693147,1048 0.693147,1050 0.693147,1059 0.693147,1061 0.693147,1082 0.693147,1083 0.693147,1102
0.693147,1110 0.693147,1111 0.693147}

Cluster 4: {0 ' Scientific Advisory Board e-print repository57',2 0.693147,10 0.693147,11 0.693147,13 0.693147,14 0.693147,19 0.693147,24 0.693147,31
0.693147,32 0.693147,36 0.693147,39 0.693147,40 0.693147,41 0.693147,43 0.693147,48 0.693147,59 0.693147,60 0.693147,61 0.693147,67 0.693147,68
0.693147,71 0.693147,72 0.693147,75 0.693147,79 0.693147,83 0.693147,89 0.693147,117 0.693147,120 0.693147,123 0.693147,136 0.693147,137
0.693147,144 0.693147,147 0.693147,152 0.693147,154 0.693147,163 0.693147,173 0.693147,175 0.693147,189 0.693147,193 0.693147,213 0.693147,223
0.693147,235 0.693147,244 0.693147,245 0.693147,248 0.693147,250 0.693147,253 0.693147,255 0.693147,260 0.693147,263 0.693147,265 0.693147,278
0.693147,279 0.693147,280 0.693147,282 0.693147,292 0.693147,294 0.693147,298 0.693147,305 0.693147,312 0.693147,322 0.693147,324 0.693147,325
0.693147,335 0.693147,341 0.693147,346 0.693147,354 0.693147,367 0.693147,368 0.693147,369 0.693147,379 0.693147,390 0.693147,392 0.693147,411
0.693147,414 0.693147,419 0.693147,430 0.693147,434 0.693147,435 0.693147,442 0.693147,443 0.693147,444 0.693147,446 0.693147,452 0.693147,459
0.693147,470 0.693147,477 0.693147,486 0.693147,492 0.693147,493 0.693147,510 0.693147,515 0.693147,525 0.693147,526 0.693147,528 0.693147,534
0.693147,541 0.693147,556 0.693147,572 0.693147,573 0.693147,578 0.693147,579 0.693147,612 0.693147,618 0.693147,626 0.693147,627 0.693147,631
0.693147,634 0.693147,652 0.693147,655 0.693147,661 0.693147,662 0.693147,663 0.693147,674 0.693147,678 0.693147,681 0.693147,684 0.693147,687
0.693147,695 0.693147,698 0.693147,710 0.693147,712 0.693147,724 0.693147,726 0.693147,752 0.693147,756 0.693147,764 0.693147,765 0.693147,776
0.693147,779 0.693147,792 0.693147,812 0.693147,816 0.693147,818 0.693147,820 0.693147,826 0.693147,830 0.693147,832 0.693147,841 0.693147,847
0.693147,853 0.693147,855 0.693147,857 0.693147,862 0.693147,883 0.693147,891 0.693147,900 0.693147,904 0.693147,907 0.693147,918 0.693147,923
0.693147,929 0.693147,934 0.693147,938 0.693147,953 0.693147,983 0.693147,985 0.693147,1004 0.693147,1008 0.693147,1010 0.693147,1020
0.693147,1024 0.693147,1025 0.693147,1035 0.693147,1042 0.693147,1048 0.693147,1050 0.693147,1059 0.693147,1082 0.693147,1083 0.693147,1093
0.693147,1102 0.693147,1106 0.693147,1110 0.693147}

Missing values globally replaced with mean/mode

Time taken to build model (full training data) : 0.06 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 2 (3%)

1 20 (29%)

2 2 (3%)

3 23 (33%)

4 22 (32%)

K-Means = 3 dla danych z filtrem WordCount:

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: dokumenty-weka.filters.unsupervised.attribute.StringToNominal-Rfirst-weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"-weka.filters.unsupervised.attribute.Remove-R2-5-weka.filters.unsupervised.attribute.StringToWordVector-Rlast-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"

Instances: 69

Attributes: 1112

[list of attributes omitted]

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 6

Within cluster sum of squared errors: 5275.052380952386

Initial starting points (random):

Cluster 0: {0 'Computer Science Subject Areas and Moderators e-40',1 1,2 1,3 1,5 1,8 1,10 1,11 1,13 1,18 1,20 1,22 1,27 1,36 1,37 1,38 1,40 1,45 1,54 1,55 1,57 1,58 1,59 1,61 1,64 1,65 1,66 1,67 1,68 1,70 1,71 1,72 1,76 1,79 1,81 1,83 1,90 1,91 1,94 1,100 1,108 1,113 1,117 1,118 1,119 1,120 1,123 1,124 1,129 1,130 1,133 1,134 1,136 1,142 1,144 1,145 1,146 1,147 1,149 1,150 1,152 1,156 1,158 1,163 1,167 1,169 1,171 1,172 1,173 1,175 1,176 1,178 1,182 1,188 1,189 1,196 1,199 1,204 1,205 1,206 1,207 1,210 1,211 1,212 1,217 1,221 1,223 1,228 1,235 1,236 1,237 1,238 1,248 1,250 1,252 1,254 1,263 1,264 1,265 1,268 1,270 1,271 1,273 1,278 1,282 1,285 1,292 1,293 1,294 1,295 1,302 1,303 1,305 1,306 1,312 1,318 1,319 1,322 1,324 1,326 1,329 1,335 1,337 1,341 1,346 1,354 1,367 1,368 1,379 1,383 1,388 1,390 1,392 1,394 1,395 1,399 1,406 1,408 1,410 1,411 1,412 1,414 1,415 1,419 1,427 1,428 1,429 1,430 1,432 1,435 1,436 1,437 1,441 1,442 1,443 1,444 1,445 1,446 1,449 1,451 1,452 1,459 1,460 1,461 1,464 1,465 1,468 1,476 1,477 1,481 1,484 1,485 1,491 1,492 1,493 1,497 1,499 1,500 1,502 1,506 1,508 1,509 1,510 1,511 1,513 1,515 1,519 1,523 1,525 1,526 1,529 1,532 1,534 1,537 1,538 1,548 1,555 1,556 1,557 1,559 1,560 1,561 1,562 1,563 1,570 1,572 1,578 1,579 1,583 1,586 1,589 1,595 1,599 1,604 1,611 1,612 1,618 1,623 1,627 1,630 1,631 1,634 1,635 1,640 1,641 1,642 1,650 1,654 1,655 1,661 1,662 1,668 1,674 1,678 1,681 1,682 1,683 1,684 1,687 1,688 1,689 1,690 1,691 1,695 1,698 1,703 1,704 1,705 1,706 1,707 1,710 1,711 1,717 1,718 1,723 1,724 1,725 1,726 1,727 1,732 1,736 1,737 1,741 1,743 1,744 1,751 1,752 1,755 1,759 1,760 1,761 1,764 1,769 1,773 1,774 1,783 1,790 1,792 1,793 1,794 1,795 1,805 1,806 1,812 1,816 1,820 1,821 1,829 1,830 1,832 1,835 1,839 1,841 1,845 1,847 1,849 1,850 1,853 1,857 1,863 1,870 1,873 1,875 1,878 1,879 1,881 1,883 1,884 1,892 1,900 1,904 1,907 1,908 1,911 1,918 1,921 1,922 1,927 1,929 1,932 1,936 1,938 1,946 1,950 1,952 1,962 1,963 1,965 1,970 1,971 1,972 1,973 1,974 1,978 1,983 1,987 1,989 1,991 1,996 1,1002 1,1004 1,1007 1,1008 1,1009 1,1011 1,1015 1,1018 1,1019 1,1024 1,1025 1,1026 1,1029 1,1030 1,1031 1,1032 1,1033 1,1034 1,1035 1,1042 1,1043 1,1045 1,1048 1,1050 1,1051 1,1059 1,1061 1,1062 1,1063 1,1065 1,1080 1,1082 1,1083 1,1093 1,1094 1,1096 1,1102 1,1105 1,1110 1}

Cluster 1: {0 'Nonlinear Sciences httpsarxivorgarchivenlin Page 25',4 1,10 1,11 1,13 1,25 1,28 1,34 1,36 1,38 1,40 1,42 1,51 1,52 1,53 1,61 1,71 1,79 1,80 1,84 1,92 1,93 1,115 1,119 1,120 1,123 1,128 1,136 1,143 1,147 1,159 1,160 1,163 1,164 1,166 1,174 1,175 1,181 1,184 1,191 1,216 1,218 1,222 1,224 1,226 1,231 1,235 1,238 1,248 1,250 1,263 1,279 1,281 1,288 1,292 1,312 1,319 1,322 1,335 1,341 1,346 1,348 1,354 1,367 1,368 1,376 1,379 1,384 1,390 1,391 1,392 1,408 1,411 1,414 1,415 1,419 1,430 1,431 1,433 1,435 1,442 1,446 1,452 1,466 1,473 1,477 1,478 1,492 1,496 1,509 1,515 1,521 1,525 1,526 1,530 1,533 1,534 1,556 1,559 1,568 1,569 1,579 1,592 1,612 1,627 1,631 1,632 1,634 1,643 1,646 1,648 1,649 1,655 1,661 1,662 1,674 1,676 1,678 1,681 1,686 1,687 1,695 1,697 1,698 1,700 1,701 1,708 1,710 1,713 1,721 1,722 1,723 1,733 1,737 1,739 1,740 1,752 1,757 1,764 1,769 1,774 1,781 1,782 1,784 1,792 1,796 1,801 1,804 1,812 1,816 1,820 1,830 1,832 1,833 1,837 1,841 1,844 1,849 1,853 1,855 1,861 1,874 1,875 1,892 1,900 1,903 1,904 1,907 1,926 1,930 1,934 1,941 1,942 1,947 1,963 1,966 1,968 1,977 1,983 1,984 1,985 1,987 1,989 1,1004 1,1008 1,1011 1,1018 1,1025 1,1035 1,1039 1,1042 1,1044 1,1046 1,1048 1,1050 1,1059 1,1075 1,1082 1,1086 1,1097 1,1102 1,1103 1,1104 1,1108 1,1110 1,1111 1}

Cluster 2: {0 'To submit an article e-print repository63',2 1,10 1,11 1,13 1,28 1,36 1,40 1,61 1,71 1,72 1,79 1,97 1,99 1,101 1,117 1,120 1,122 1,123 1,136 1,137 1,144 1,147 1,149 1,163 1,170 1,173 1,175 1,202 1,215 1,217 1,223 1,232 1,235 1,248 1,249 1,250 1,259 1,263 1,282 1,286 1,292 1,297 1,308 1,309 1,310 1,312 1,314 1,321 1,322 1,324 1,328 1,329 1,332 1,333 1,335 1,336 1,340 1,341 1,345 1,346 1,350 1,354 1,356 1,358 1,361 1,362 1,367 1,368 1,379 1,380 1,383 1,387 1,390 1,392 1,400 1,402 1,408 1,410 1,411 1,414 1,418 1,419 1,421 1,422 1,423 1,424 1,430 1,433 1,435 1,439 1,440 1,442 1,443 1,444 1,446 1,447 1,449 1,452 1,453 1,456 1,457 1,459 1,460 1,461 1,463 1,465 1,467 1,476 1,477 1,479 1,481 1,482 1,483 1,487 1,488 1,489 1,492 1,493 1,513 1,515 1,525 1,526 1,531 1,534 1,537 1,539 1,540 1,556 1,561 1,562 1,566 1,567 1,570 1,572 1,573 1,579 1,591 1,595 1,606 1,609 1,612 1,614 1,615 1,616 1,626 1,627 1,629 1,631 1,633 1,634 1,639 1,641 1,645 1,655 1,660 1,661 1,662 1,663 1,674 1,678 1,681 1,682 1,684 1,687 1,688 1,691 1,695 1,698 1,699 1,710 1,711 1,712 1,720 1,724 1,728 1,730 1,735 1,738 1,741 1,743 1,745 1,752 1,759 1,764 1,768 1,779 1,784 1,787 1,789 1,791 1,792 1,793 1,796 1,799 1,806 1,807 1,812 1,816 1,820 1,822 1,824 1,825 1,830 1,831 1,832 1,834 1,835 1,837 1,840 1,841 1,853 1,857 1,861 1,865 1,872 1,873 1,875 1,881 1,882 1,884 1,885 1,898 1,899 1,900 1,904 1,906 1,907 1,908 1,916 1,917 1,918 1,921 1,928 1,929 1,934 1,935 1,945 1,946 1,948 1,952 1,962 1,968 1,970 1,975 1,976 1,983 1,992 1,993 1,995 1,996 1,1002 1,1004 1,1008 1,1010 1,1018 1,1020 1,1024 1,1025 1,1026 1,1027 1,1028 1,1031 1,1032 1,1034 1,1035 1,1038 1,1039 1,1041 1,1042 1,1048 1,1050 1,1053 1,1055 1,1059 1,1061 1,1062 1,1063 1,1064 1,1067 1,1069 1,1071 1,1072 1,1080 1,1082 1,1083 1,1093 1,1094 1,1096 1,1101 1,1102 1,1105 1,1110 1,1111 1}

Missing values globally replaced with mean/mode

Time taken to build model (full training data) : 0.07 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 2 (3%)

1 42 (61%)

2 25 (36%)

K-Means = 5 dla danych z filtrem WordCount:

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: dokumenty-weka.filters.unsupervised.attribute.StringToNominal-Rfirst-weka.filters.unsupervised.attribute.StringToWordVector-R2-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"-weka.filters.unsupervised.attribute.Remove-R2-5-weka.filters.unsupervised.attribute.StringToWordVector-Rlast-W1000-prune-rate-1.0-N0-stemmerweka.core.stemmers.NullStemmer-stopwords-handlerweka.core.stopwords.Null-M1-tokenizerweka.core.tokenizers.WordTokenizer -delimiters " \r\n\t.,;:\'\"()?!"

Instances: 69

Attributes: 1112

[list of attributes omitted]

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 3

Within cluster sum of squared errors: 2560.4531620553344

Initial starting points (random):

Cluster 0: {0 'Computer Science Subject Areas and Moderators e-40',1 1,2 1,3 1,5 1,8 1,10 1,11 1,13 1,18 1,20 1,22 1,27 1,36 1,37 1,38 1,40 1,45 1,54 1,55 1,57 1,58 1,59 1,61 1,64 1,65 1,66 1,67 1,68 1,70 1,71 1,72 1,76 1,79 1,81 1,83 1,90 1,91 1,94 1,100 1,108 1,113 1,117 1,118 1,119 1,120 1,123 1,124 1,129 1,130 1,133 1,134 1,136 1,142 1,144 1,145 1,146 1,147 1,149 1,150 1,152 1,156 1,158 1,163 1,167 1,169 1,171 1,172 1,173 1,175 1,176 1,178 1,182 1,188 1,189 1,196 1,199 1,204 1,205 1,206 1,207 1,210 1,211 1,212 1,217 1,221 1,223 1,228 1,235 1,236 1,237 1,238 1,248 1,250 1,252 1,254 1,263 1,264 1,265 1,268 1,270 1,271 1,273 1,278 1,282 1,285 1,292 1,293 1,294 1,295 1,302 1,303 1,305 1,306 1,312 1,318 1,319 1,322 1,324 1,326 1,329 1,335 1,337 1,341 1,346 1,354 1,367 1,368 1,379 1,383 1,388 1,390 1,392 1,394 1,395 1,399 1,406 1,408 1,410 1,411 1,412 1,414 1,415 1,419 1,427 1,428 1,429 1,430 1,432 1,435 1,436 1,437 1,441 1,442 1,443 1,444 1,445 1,446 1,449 1,451 1,452 1,459 1,460 1,461 1,464 1,465 1,468 1,476 1,477 1,481 1,484 1,485 1,491 1,492 1,493 1,497 1,499 1,500 1,502 1,506 1,508 1,509 1,510 1,511 1,513 1,515 1,519 1,523 1,525 1,526 1,529 1,532 1,534 1,537 1,538 1,548 1,555 1,556 1,557 1,559 1,560 1,561 1,562 1,563 1,570 1,572 1,578 1,579 1,583 1,586 1,589 1,595 1,599 1,604 1,611 1,612 1,618 1,623 1,627 1,630 1,631 1,634 1,635 1,640 1,641 1,642 1,650 1,654 1,655 1,661 1,662 1,668 1,674 1,678 1,681 1,682 1,683 1,684 1,687 1,688 1,689 1,690 1,691 1,695 1,698 1,703 1,704 1,705 1,706 1,707 1,710 1,711 1,717 1,718 1,723 1,724 1,725 1,726 1,727 1,732 1,736 1,737 1,741 1,743 1,744 1,751 1,752 1,755 1,759 1,760 1,761 1,764 1,769 1,773 1,774 1,783 1,790 1,792 1,793 1,794 1,795 1,805 1,806 1,812 1,816 1,820 1,821 1,829 1,830 1,832 1,835 1,839 1,841 1,845 1,847 1,849 1,850 1,853 1,857 1,863 1,870 1,873 1,875 1,878 1,879 1,881 1,883 1,884 1,892 1,900 1,904 1,907 1,908 1,911 1,918 1,921 1,922 1,927 1,929 1,932 1,936 1,938 1,946 1,950 1,952 1,962 1,963 1,965 1,970 1,971 1,972 1,973 1,974 1,978 1,983 1,987 1,989 1,991 1,996 1,1002 1,1004 1,1007 1,1008 1,1009 1,1011 1,1015 1,1018 1,1019 1,1024 1,1025 1,1026 1,1029 1,1030 1,1031 1,1032 1,1033 1,1034 1,1035 1,1042 1,1043 1,1045 1,1048 1,1050 1,1051 1,1059 1,1061 1,1062 1,1063 1,1065 1,1080 1,1082 1,1083 1,1093 1,1094 1,1096 1,1102 1,1105 1,1110 1}

Cluster 1: {0 'Nonlinear Sciences httpsarxivorgarchivenlin Page 25',4 1,10 1,11 1,13 1,25 1,28 1,34 1,36 1,38 1,40 1,42 1,51 1,52 1,53 1,61 1,71 1,79 1,80 1,84 1,92 1,93 1,115 1,119 1,120 1,123 1,128 1,136 1,143 1,147 1,159 1,160 1,163 1,164 1,166 1,174 1,175 1,181 1,184 1,191 1,216 1,218 1,222 1,224 1,226 1,231 1,235 1,238 1,248 1,250 1,263 1,279 1,281 1,288 1,292 1,312 1,319 1,322 1,335 1,341 1,346 1,348 1,354 1,367 1,368 1,376 1,379 1,384 1,390 1,391 1,392 1,408 1,411 1,414 1,415 1,419 1,430 1,431 1,433 1,435 1,442 1,446 1,452 1,466 1,473 1,477 1,478 1,492 1,496 1,509 1,515 1,521 1,525 1,526 1,530 1,533 1,534 1,556 1,559 1,568 1,569 1,579 1,592 1,612 1,627 1,631 1,632 1,634 1,643 1,646 1,648 1,649 1,655 1,661 1,662 1,674 1,676 1,678 1,681 1,686 1,687 1,695 1,697 1,698 1,700 1,701 1,708 1,710 1,713 1,721 1,722 1,723 1,733 1,737 1,739 1,740 1,752 1,757 1,764 1,769 1,774 1,781 1,782 1,784 1,792 1,796 1,801 1,804 1,812 1,816 1,820 1,830 1,832 1,833 1,837 1,841 1,844 1,849 1,853 1,855 1,861 1,874 1,875 1,892 1,900 1,903 1,904 1,907 1,926 1,930 1,934

1,941 1,942 1,947 1,963 1,966 1,968 1,977 1,983 1,984 1,985 1,987 1,989 1,1004 1,1008 1,1011 1,1018 1,1025 1,1035 1,1039 1,1042 1,1044 1,1046 1,1048 1,1050 1,1059 1,1075 1,1082 1,1086 1,1097 1,1102 1,1103 1,1104 1,1108 1,1110 1,1111 1}

Cluster 2: {0 'To submit an article e-print repository63',2 1,10 1,11 1,13 1,28 1,36 1,40 1,61 1,71 1,72 1,79 1,97 1,99 1,101 1,117 1,120 1,122 1,123 1,136 1,137 1,144 1,147 1,149 1,163 1,170 1,173 1,175 1,202 1,215 1,217 1,223 1,232 1,235 1,248 1,249 1,250 1,259 1,263 1,282 1,286 1,292 1,297 1,308 1,309 1,310 1,312 1,314 1,321 1,322 1,324 1,328 1,329 1,332 1,333 1,335 1,336 1,340 1,341 1,345 1,346 1,350 1,354 1,356 1,358 1,361 1,362 1,367 1,368 1,379 1,380 1,383 1,387 1,390 1,392 1,400 1,402 1,408 1,410 1,411 1,414 1,418 1,419 1,421 1,422 1,423 1,424 1,430 1,433 1,435 1,439 1,440 1,442 1,443 1,444 1,446 1,447 1,449 1,452 1,453 1,456 1,457 1,459 1,460 1,461 1,463 1,465 1,467 1,476 1,477 1,479 1,481 1,482 1,483 1,487 1,488 1,489 1,492 1,493 1,513 1,515 1,525 1,526 1,531 1,534 1,537 1,539 1,540 1,556 1,561 1,562 1,566 1,567 1,570 1,572 1,573 1,579 1,591 1,595 1,606 1,609 1,612 1,614 1,615 1,616 1,626 1,627 1,629 1,631 1,633 1,634 1,639 1,641 1,645 1,655 1,660 1,661 1,662 1,663 1,674 1,678 1,681 1,682 1,684 1,687 1,688 1,691 1,695 1,698 1,699 1,710 1,711 1,712 1,720 1,724 1,728 1,730 1,735 1,738 1,741 1,743 1,745 1,752 1,759 1,764 1,768 1,779 1,784 1,787 1,789 1,791 1,792 1,793 1,796 1,799 1,806 1,807 1,812 1,816 1,820 1,822 1,824 1,825 1,830 1,831 1,832 1,834 1,835 1,837 1,840 1,841 1,853 1,857 1,861 1,865 1,872 1,873 1,875 1,881 1,882 1,884 1,885 1,898 1,899 1,900 1,904 1,906 1,907 1,908 1,916 1,917 1,918 1,921 1,928 1,929 1,934 1,935 1,945 1,946 1,948 1,952 1,962 1,968 1,970 1,975 1,976 1,983 1,992 1,993 1,995 1,996 1,1002 1,1004 1,1008 1,1010 1,1018 1,1020 1,1024 1,1025 1,1026 1,1027 1,1028 1,1031 1,1032 1,1034 1,1035 1,1038 1,1039 1,1041 1,1042 1,1048 1,1050 1,1053 1,1055 1,1059 1,1061 1,1062 1,1063 1,1064 1,1067 1,1069 1,1071 1,1072 1,1080 1,1082 1,1083 1,1093 1,1094 1,1096 1,1101 1,1102 1,1105 1,1110 1,1111 1}

Cluster 3: {0 'Contacting e-print repository httpsarxivorghelp68',2 1,10 1,11 1,13 1,36 1,40 1,61 1,71 1,72 1,79 1,97 1,99 1,114 1,117 1,120 1,122 1,123 1,136 1,137 1,144 1,147 1,163 1,173 1,175 1,194 1,202 1,217 1,223 1,235 1,247 1,248 1,250 1,263 1,282 1,287 1,292 1,312 1,322 1,335 1,336 1,341 1,346 1,354 1,367 1,368 1,379 1,380 1,390 1,392 1,399 1,408 1,411 1,414 1,419 1,430 1,435 1,442 1,443 1,444 1,446 1,449 1,452 1,454 1,459 1,461 1,479 1,482 1,492 1,493 1,515 1,522 1,525 1,526 1,546 1,550 1,556 1,570 1,572 1,579 1,584 1,595 1,612 1,617 1,627 1,631 1,634 1,639 1,641 1,655 1,661 1,662 1,663 1,674 1,677 1,678 1,681 1,684 1,687 1,691 1,695 1,698 1,699 1,710 1,724 1,752 1,759 1,764 1,766 1,767 1,776 1,779 1,792 1,804 1,812 1,816 1,820 1,830 1,832 1,835 1,837 1,841 1,843 1,853 1,857 1,860 1,869 1,894 1,898 1,900 1,901 1,904 1,907 1,911 1,918 1,920 1,924 1,929 1,934 1,973 1,983 1,991 1,992 1,1004 1,1008 1,1012 1,1020 1,1024 1,1025 1,1031 1,1035 1,1038 1,1039 1,1042 1,1048 1,1050 1,1059 1,1061 1,1082 1,1083 1,1102 1,1110 1,1111 1}

Cluster 4: {0 ' Scientific Advisory Board e-print repository57',2 1,10 1,11 1,13 1,14 1,19 1,24 1,31 1,32 1,36 1,39 1,40 1,41 1,43 1,48 1,59 1,60 1,61 1,67 1,68 1,71 1,72 1,75 1,79 1,83 1,89 1,117 1,120 1,123 1,136 1,137 1,144 1,147 1,152 1,154 1,163 1,173 1,175 1,189 1,193 1,213 1,223 1,235 1,244 1,245 1,248 1,250 1,253 1,255 1,260 1,263 1,265 1,278 1,279 1,280 1,282 1,292 1,294 1,298 1,305 1,312 1,322 1,324 1,325 1,335 1,341 1,346 1,354 1,367 1,368 1,369 1,379 1,390 1,392 1,411 1,414 1,419 1,430 1,434 1,435 1,442 1,443 1,444 1,446 1,452 1,459 1,470 1,477 1,486 1,492 1,493 1,510 1,515 1,525 1,526 1,528 1,534 1,541 1,556 1,572 1,573 1,578 1,579 1,612 1,618 1,626 1,627 1,631 1,634 1,652 1,655 1,661 1,662 1,663 1,674 1,678 1,681 1,684 1,687 1,695 1,698 1,710 1,712 1,724 1,726 1,752 1,756 1,764 1,765 1,776 1,779 1,792 1,812 1,816 1,818 1,820 1,826 1,830 1,832 1,841 1,847 1,853 1,855 1,857 1,862 1,883 1,891 1,900 1,904 1,907 1,918 1,923 1,929 1,934 1,938 1,953 1,983 1,985 1,1004 1,1008 1,1010 1,1020 1,1024 1,1025 1,1035 1,1042 1,1048 1,1050 1,1059 1,1082 1,1083 1,1093 1,1102 1,1106 1,1110 1}

Missing values globally replaced with mean/mode

Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 2 (3%)

1 20 (29%)

2 2 (3%)

3 23 (33%)

4 22 (32%)

Wnioski końcowe:

Program WebSphinx jest bardzo praktycznym narzędziem które pozwala na pozyskanie danych internetowych. Aplikacja w sposób intuicyjny tworzy nasz dokument na podstawie podanej strony internetowej. Zapewnia ona wsparcie w przypadku potrzeby analizy strony pod kątem zawartości oraz jej dokładnego występowania. Dane pozyskane w programie WebSphinx należy następnie przekonwertować w plik textowy oraz końcowy plik typu .arff w celu jego dalszej analizy programem "Weka". Aplikacja "Weka" umożliwia nam zamianę atrybutu zawartości w wektor słów oraz jego dalsze filtrowanie np. pod kątem:

- bezwzględna wartość wystąpień danych słów w zbiorze dokumentów (word-count)

- znormalizowana wartość wystąpień danych słów w zbiorze dokumentów i w kontekście wystąpienia tych słów (TFTransform, IDFTransform)
- Binaryzacja wyników oraz ich wizualizacja

Klastrowanie danych nie przefiltrowanych uzyskało wyniki dla k-means=3:

- klaster nr 1: 3%
- klaster nr 2: 61%
- klaster nr 3: 36%

dla k-means=5 klastrów wyniki:

- klaster nr 1: 3%
- klaster nr 2: 29%
- klaster nr 3: 3%
- klaster nr 4: 33%
- klaster nr 5: 32%

Oznacza to że zawartość strony generalnie dzieli się na klastry ze słowami rzadko występującymi we wszystkich dokumentach (3%) oraz klastry ze słowami występującymi często (29-61%). Wynik wskazuje więc na bardzo unikatową zawartość słów w części dokumentów. Pokrywa się to częściowo z analizą danych nieprzefiltrowanych - w tabeli widoczne są słowa występujące tylko w pojedynczym lub kilku dokumentach. Większość słów występuje natomiast dosyć często w pozostałych dokumentach.

Klastrowanie danych przefiltrowanych filtrem NumericToBinary uzyskało wyniki dla k-means=3:

- klaster nr 1: 3%
- klaster nr 2: 86%
- klaster nr 3: 12%

Klastrowanie danych przefiltrowanych filtrem NumericToBinary uzyskało wyniki dla k-means=5:

- klaster nr 1: 3%
- klaster nr 2: 29%
- klaster nr 3: 4%
- klaster nr 4: 33%
- klaster nr 5: 30%

Uzyskane wyniki przypominają te które zostały uzyskane bez filtrowania, szczególnie dla przypadku k-means=5. W przypadku k-means=3 widać przewagę jednego klastra 86% nad pozostałymi.

Klastrowanie danych przefiltrowanych filtrem IDFTransform uzyskało wyniki dla k-means=3:

- klaster nr 1: 3%
- klaster nr 2: 61%
- klaster nr 3: 36%

Klastrowanie danych przefiltrowanych filtrem IDFTransform uzyskało wyniki dla k-means=5:

- klaster nr 1: 3%
- klaster nr 2: 29%
- klaster nr 3: 3%
- klaster nr 4: 33%
- klaster nr 5: 32%

Klastrowanie danych przefiltrowanych filtremTFTransform uzyskało wyniki dla k-means=3:

- klaster nr 1: 3%
- klaster nr 2: 61%

- klaster nr 3: 36%

Klastrowanie danych przefiltrowanych filtrem TFTransform uzyskało wyniki dla k-means=5:

- klaster nr 1: 3%
- klaster nr 2: 29%
- klaster nr 3: 3%
- klaster nr 4: 33%
- klaster nr 5: 32%

Klastrowanie danych przefiltrowanych filtrem WordCount uzyskało wyniki dla k-means=3:

- klaster nr 1: 3%
- klaster nr 2: 61%
- klaster nr 3: 36%

Klastrowanie danych przefiltrowanych filtrem WordCount uzyskało wyniki dla k-means=5:

- klaster nr 1: 3%
- klaster nr 2: 29%
- klaster nr 3: 3%
- klaster nr 4: 33%
- klaster nr 5: 32%

Jak widać powyżej wyniki klastrowania dla filtrów IDFTransform, TFTransform oraz WordCount są zbieżne. Wynika to z charakteru danych, Filtry IDFTransform oraz TFTransform normalizują wartości. nie zmieniają ani nie wpływają na względną ilość wystąpień słów w danych dokumentach.