DAT630
Text Classification and Clustering

Search Engines, Chapters 4, 9

27/09/2016

Krisztian Balog | University of Stavanger

So far

- We worked with **record data**
  - Each record is described by a set of attributes
  - Often, we prefer to work with attributes of the same type
    - E.g., convert everything to categorical for Decision Trees, convert everything to numerical for SVM
- Handful of attributes (low dimensionality)
- Straighforward to compare records
  - E.g., Eucledian distance

Document Data

- Records (or objects) are **documents**
  - Web pages, emails, books, text messages, tweets, Facebook pages, MS Office documents, etc.
- Core ingredient for classification and clustering: **measuring similarity**
- Questions when working with documents:
  - How to represent documents?
  - How to measure the similarity between documents?

Issues

- Text is noisy
  - Variations in spelling
  - Morphological variations. E.g.,
    - car, cars, car's
    - take, took, taking, taken, …
- Text is ambiguous
  - Many different ways to express the same meaning

Representing Documents

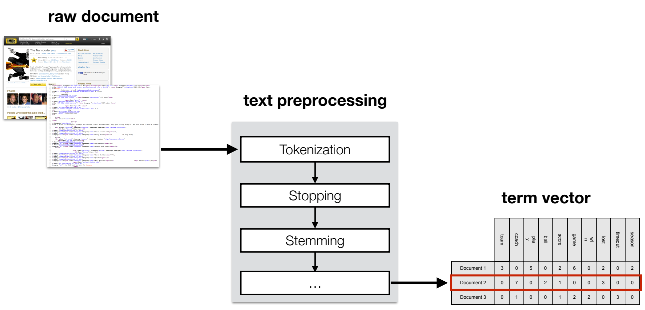
- Documents are represented as **term vectors**
  - Each term is a component (attribute) of the vector
  - Values correspond to the number of times the term appears in the document

	team	coach	pla y	ball	score	game	wi n	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Term-document (or document-term) matrix

Text Preprocessing

Preprocessing Pipeline



Tokenization

- Parsing a string into individual words (tokens)
- Splitting is usually done along white spaces, punctuation marks, or other types of content delimiters (e.g., HTML markup)
- Sounds easy, but can be surprisingly complex, even for English
  - Even worse for many other languages

Tokenization Issues

- Apostrophes can be a part of a word, a part of a possessive, or just a mistake
  - rosie o'donnell, can't, 80's, 1890's, men's straw hats, master's degree, …
- Capitalized words can have different meaning from lower case words
  - Bush, Apple
- Special characters are an important part of tags, URLs, email addresses, etc.
  - C++, C#, …

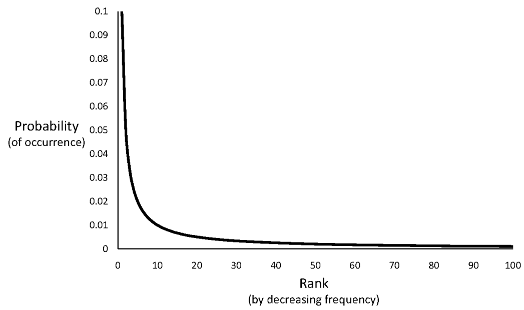
Tokenization Issues

- Numbers can be important, including decimals
  - nokia 3250, top 10 courses, united 93, quicktime 6.5 pro, 92.3 the beat, 288358
- Periods can occur in numbers, abbreviations, URLs, ends of sentences, and other situations
  - I.B.M., Ph.D., www.uis.no, F.E.A.R.

Common Practice

- First pass is focused on identifying markup or tags; second pass is done on the appropriate parts of the document structure
- Treat hyphens, apostrophes, periods, etc. like spaces
- Ignore capitalization
- Index even single characters
  - o'connor => o connor

Text Statistics



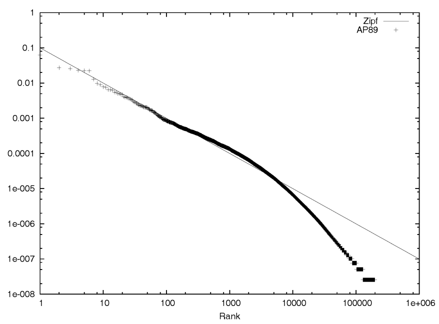
Top-50 words from AP89

Word	*Freq.*	*r*	P_r(%)	$r.P_r$	*Word*	*Freq*	*r*	P_r(%)	$r.P_r$
the	2,420,778	1	6.49	0.065	has	136,007	26	0.37	0.095
of	1,045,733	2	2.80	0.056	are	130,322	27	0.35	0.094
to	968,882	3	2.60	0.078	not	127,493	28	0.34	0.096
a	892,429	4	2.39	0.096	who	116,364	29	0.31	0.090
and	865,644	5	2.32	0.120	they	111,024	30	0.30	0.089
in	847,825	6	2.27	0.140	its	111,021	31	0.30	0.092
said	504,593	7	1.35	0.095	had	103,943	32	0.28	0.089
for	363,865	8	0.98	0.078	will	102,949	33	0.28	0.091
that	347,072	9	0.93	0.084	would	99,503	34	0.27	0.091
was	293,027	10	0.79	0.079	about	92,983	35	0.25	0.087
on	291,947	11	0.78	0.086	i	92,005	36	0.25	0.089
he	250,919	12	0.67	0.081	been	88,786	37	0.24	0.088
is	245,843	13	0.65	0.086	this	87,286	38	0.23	0.089
with	223,846	14	0.60	0.084	their	84,638	39	0.23	0.089
at	210,064	15	0.56	0.085	new	83,449	40	0.22	0.090
by	209,586	16	0.56	0.090	or	81,796	41	0.22	0.090
it	195,621	17	0.52	0.089	which	80,385	42	0.22	0.091
from	189,451	18	0.51	0.091	we	80,245	43	0.22	0.093
as	181,714	19	0.49	0.093	more	76,388	44	0.21	0.090
be	157,300	20	0.42	0.084	after	75,165	45	0.20	0.091
were	153,913	21	0.41	0.087	us	72,045	46	0.19	0.089
an	152,576	22	0.41	0.090	percent	71,956	47	0.19	0.091
have	149,749	23	0.40	0.092	up	71,082	48	0.19	0.092
his	142,285	24	0.38	0.092	one	70,266	49	0.19	0.092
but	140,880	25	0.38	0.094	people	68,988	50	0.19	0.093

Zipf's Law

- Distribution of word frequencies is very *skewed*
  - A few words occur very often, many words hardly ever occur
  - E.g., two most common words ("the", "of") make up about 10% of all word occurrences in text documents
- Zipf's law:
  - Frequency of an item or event is inversely proportional to its frequency rank
  - Rank (r) of a word times its frequency (f) is approximately a constant (k): r*f~k

Zip's law for AP89



Stopword Removal

- Function words that have little meaning apart from other words: the, a, an, that, those, …
- These are considered *stopwords* and are removed
- A stopwords list can be constructed by taking the top n (e.g., 50) most common words in a collection

Stopword Removal

a	as	by	into	not	such	then	this	with
an	at	for	is	of	that	there	to	
and	be	if	it	on	the	these	was	
are	but	in	no	or	their	they	will	

Table 2: Standard English stopwords list.

- There are problematic cases…

"to be or not to be"

Stopword Removal

- Lists are customized for applications, domains, and even parts of documents
  - E.g., "click" is a good stopword for anchor text

Stemming

- Reduce the different forms of a word that occur to a common *stem*
  - inflectional (plurals, tenses)
  - derivational (making verbs nouns etc.)
- In most cases, these have the same or very similar meanings
- Two basic types of stemmers
  - Algorithmic
  - Dictionary-based

Stemming

- **Suffix-s stemmer**
  - Assumes that any word ending with an s is plural
    - cakes => cake, dogs =>dog
  - Cannot detect many plural relationships (false negative)
    - centuries => century
  - In rare cases it detects a relationship where it does not exist (false positive)
    - is => i

Stemming

- **Porter stemmer**
  - Most popular algorithmic stemmer
  - Consists of 5 steps, each step containing a set of rules for removing suffixes
  - Produces stems not words
  - Makes a number of errors and difficult to modify

Porter Stemmer

- Example step (1 of 5)

Step 1a:

- Replace ***sses*** by ***ss*** (e.g., stresses → stress).
- Delete ***s*** if the preceding word part contains a vowel not immediately before the ***s*** (e.g., gaps → gap but gas → gas).
- Replace ***ied*** or ***ies*** by ***i*** if preceded by more than one letter, otherwise by ***ie*** (e.g., ties → tie, cries → cri).
- If suffix is ***us*** or ***ss*** do nothing (e.g., stress → stress).

Step 1b:

- Replace ***eed***, ***eedly*** by ***ee*** if it is in the part of the word after the first non-vowel following a vowel (e.g., agreed → agree, feed → feed).
- Delete ***ed***, ***edly***, ***ing***, ***ingly*** if the preceding word part contains a vowel, and then if the word ends in ***at***, ***bl***, or ***iz*** add ***e*** (e.g., fished → fish, pirating → pirate), or if the word ends with a double letter that is not ***ll***, ***ss*** or ***zz***, remove the last letter (e.g., falling→ fall, dripping → drip), or if the word is short, add ***e*** (e.g., hoping → hope).
- Whew!

Porter Stemmer

should not have the same stem

should have the same stem

False positives	*False negatives*
organization/organ	european/europe
generalization/generic	cylinder/cylindrical
numerical/numerous	matrices/matrix
policy/police	urgency/urgent
university/universe	create/creation
addition/additive	analysis/analyses
negligible/negligent	useful/usefully
execute/executive	noise/noisy
past/paste	decompose/decomposition
ignore/ignorant	sparse/sparsity
special/specialized	resolve/resolution
head/heading	triangle/triangular

Stemming

- **Krovetz stemmer**
  - Hybrid algorithmic-dictionary
  - Word checked in dictionary
    - If present, either left alone or replaced with exception stems
    - If not present, word is checked for suffixes that could be removed
  - After removal, dictionary is checked again
  - Produces words not stems

Stemmer Comparison

Original text

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales

Porter stemmer

market strateg carr compan agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale stimul demand price cut volum sale

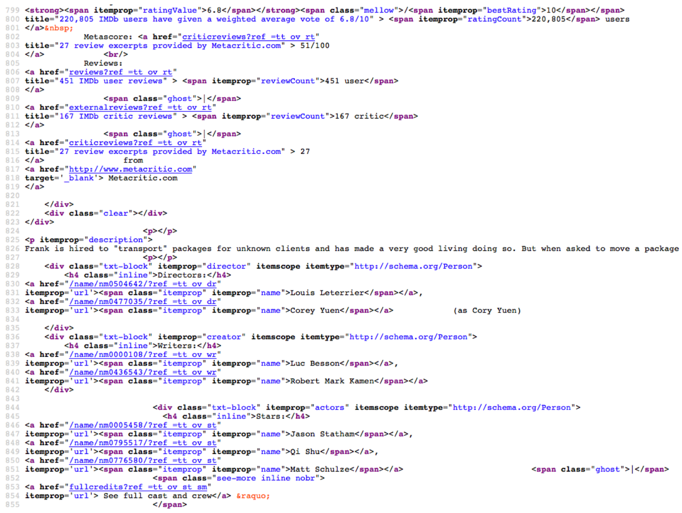
Krovetz stemmer

marketing strategy carry company agriculture chemical report prediction market share chemical report market statistic agrochemic pesticide herbicide fungicide insecticide fertilizer predict sale stimulate demand price cut volume sale

Stemming

- Generally a small (but significant) effectiveness improvement for English
- Can be crucial for some languages (e.g., Arabic, Russian)

Example



```html
<strong><span itemprop="ratingValue">6.8</span></strong><span class="mellow">/<span itemprop="bestRating">10</span></span>
title="220,805 IMDb users have given a weighted average vote of 6.8/10" > <span itemprop="ratingCount">220,805</span> users
</a> 
            Metascore: <a href="criticreviews?ref_=tt_ov_rt"
title="27 review excerpts provided by Metacritic.com" > 51/100
</a>            <br/>
            Reviews:
<a href="reviews?ref_=tt_ov_rt"
title="451 IMDb user reviews" > <span itemprop="reviewCount">451 user</span>
</a>
                <span class="ghost">|</span>
<a href="externalreviews?ref_=tt_ov_rt"
title="167 IMDb critic reviews" > <span itemprop="reviewCount">167 critic</span>
</a>
                <span class="ghost">|</span>
<a href="criticreviews?ref_=tt_ov_rt"
title="27 review excerpts provided by Metacritic.com" > 27
</a>                from
<a href="http://www.metacritic.com"
target='_blank'> Metacritic.com
</a>

    </div>
    <div class="clear"></div>
</div>
                        <p></p>
<p itemprop="description">
Frank is hired to "transport" packages for unknown clients and has made a very good living doing so. But when asked to move a package
                        <p></p>
    <div class="txt-block" itemprop="director" itemscope itemtype="http://schema.org/Person">
        <h4 class="inline">Directors:</h4>
<a href="/name/nm0504642/?ref_=tt_ov_dr"
itemprop='url'><span class="itemprop" itemprop="name">Louis Leterrier</span></a>,
<a href="/name/nm0477035/?ref_=tt_ov_dr"
itemprop='url'><span class="itemprop" itemprop="name">Corey Yuen</span></a>            (as Cory Yuen)

    </div>
    <div class="txt-block" itemprop="creator" itemscope itemtype="http://schema.org/Person">
        <h4 class="inline">Writers:</h4>
<a href="/name/nm0000108/?ref_=tt_ov_wr"
itemprop='url'><span class="itemprop" itemprop="name">Luc Besson</span></a>,
<a href="/name/nm0436543/?ref_=tt_ov_wr"
itemprop='url'><span class="itemprop" itemprop="name">Robert Mark Kamen</span></a>
    </div>

                          <div class="txt-block" itemprop="actors" itemscope itemtype="http://schema.org/Person">
                            <h4 class="inline">Stars:</h4>
<a href="/name/nm0005458/?ref_=tt_ov_st"
itemprop='url'><span class="itemprop" itemprop="name">Jason Statham</span></a>,
<a href="/name/nm0795517/?ref_=tt_ov_st"
itemprop='url'><span class="itemprop" itemprop="name">Qi Shu</span></a>,
<a href="/name/nm0776580/?ref_=tt_ov_st"
itemprop='url'><span class="itemprop" itemprop="name">Matt Schulze</span></a>            <span class="ghost">|</span>
                          <span class="see-more inline nobr">
<a href="fullcredits?ref_=tt_ov_st_sm"
itemprop='url'> See full cast and crew</a> &raquo;
                          </span>
```

First pass extraction

The Transporter (2002)
PG-13 92 min Action, Crime, Thriller 11 October 2002 (USA)

Frank is hired to "transport" packages for unknown clients and has made a very good living doing so. But when asked to move a package that begins moving, complications arise.

Tokenization

the transporter 2002
pg 13 92 min action crime thriller 11 october 2002 usa

frank is hired to transport packages for unknown clients and has made a very good living doing so but when asked to move a package that begins moving complications arise

Stopwords removal

the transporter 2002
pg 13 92 min action crime thriller 11 october 2002 usa

frank is hired to transport packages for unknown clients and has made a very good living doing so but when asked to move a package that begins moving complications arise

transporter 2002
pg 13 92 min action crime thriller 11 october 2002 usa

frank hired transport packages unknown clients has made very good living doing so when asked move package begins moving complications arise

Stemming (Porter stemmer)

transporter 2002
pg 13 92 min action crime thriller 11 october 2002 usa

frank hired transport packages unknown clients has made very good living doing so when asked move package begins moving complications arise

transport 2002
pg 13 92 min action crime thriller 11 octob 2002 usa

frank hire transport packag unknown client ha made veri good live do so when ask move packag begin move complic aris

Exercise

- Task 1

Bag-of-words Model

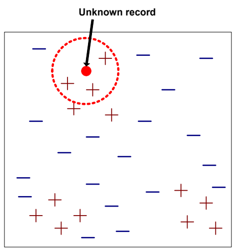
- Simplifying representation
- Text (document) is represented as the **bag** (multiset) of its **words**
- Disregards word ordering, but keeps multiplicity
  - I.e., positional independence assumption is made

Text Classification

K-Nearest Neighbor

K-Nearest Neighbor (KNN)

- Instance-based classifier that uses the K "closest" points (nearest neighbors) for performing classification



KNN for Text Classification

- Represent documents as points (vectors)
- Define a **similarity measure** for pairwise documents
- Select the value of K
- Choose a voting scheme (e.g., majority vote) to determine the class label of an unseen document

Similarity Measures

- T_1 and T_2 are the set of terms in d_1 and d_2
- **Number of overlapping words** $|T_1 \cap T_2|$
  - Fails to account for document size
    - Long documents will have more overlapping words than short ones
- **Jaccard similarity** $\frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$
  - Produces a number between 0 and 1
  - Considers only presence/absence of terms, does not take into account actual term frequencies

Similarity Measures

- **Cosine similarity**
  - $\vec{d_1}$ and $\vec{d_2}$ are document vectors with term freqs.

$$cos(\vec{d_1}, \vec{d_2}) = \frac{\boxed{\vec{d_1} \cdot \vec{d_2}}}{\boxed{||\vec{d_1}||}\boxed{||\vec{d_2}||}}$$

$$\vec{d_1} \cdot \vec{d_2} \longrightarrow \sum_t n(t, d_1) n(t, d_2)$$

$$||\vec{d_1}|| \rightarrow \sqrt{\sum_t n(t, d_1)^2} \qquad ||\vec{d_2}|| \rightarrow \sqrt{\sum_t n(t, d_2)^2}$$

Example

	term 1	term 2	term 3	term 4	term 5
doc 1	1	0	1	0	3
doc 2	0	2	4	0	1

$$cos(\vec{d_1}, \vec{d_2}) = \frac{\boxed{\vec{d_1} \cdot \vec{d_2}}}{\boxed{||\vec{d_1}||}\boxed{||\vec{d_2}||}}$$

$$\vec{d_1} \cdot \vec{d_2} \longrightarrow \sum_t n(t, d_1) n(t, d_2)$$

$$||\vec{d_1}|| \rightarrow \sqrt{\sum_t n(t, d_1)^2} \qquad ||\vec{d_2}|| \rightarrow \sqrt{\sum_t n(t, d_2)^2}$$

Example

	term 1	term 2	term 3	term 4	term 5
doc 1	1	0	1	0	3
doc 2	0	2	4	0	1

$$cos(\vec{d_1}, \vec{d_2}) = \frac{\boxed{\vec{d_1} \cdot \vec{d_2}}}{\boxed{||\vec{d_1}||}\boxed{||\vec{d_2}||}}$$

7/(3.31*4.58)=**0.46**

$$\vec{d_1} \cdot \vec{d_2} \longrightarrow \sum_t n(t, d_1) n(t, d_2)$$

1*0+0*2+1*4+0*0+3*1=7

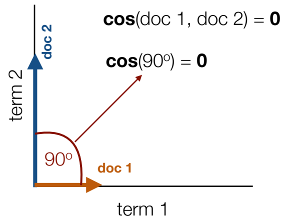
$$||\vec{d_1}|| = \sqrt{\sum_t n(t, d_1)^2} \qquad ||\vec{d_2}|| = \sqrt{\sum_t n(t, d_2)^2}$$

sqrt($1^2+0^2+1^2+0^2+3^2$)=sqrt(11)=**3.31**

sqrt($0^2+2^2+4^2+0^2+1^2$)=sqrt(21)=**4.58**

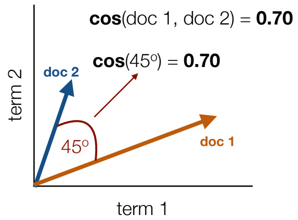
Geometric Interpretation

	term 1	term 2
doc 1	1	0
doc 2	0	2



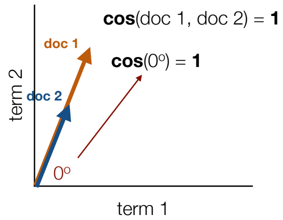
Geometric Interpretation

	term 1	term 2
doc 1	4	2
doc 2	1	3



Geometric Interpretation

	term 1	term 2
doc 1	1	2
doc 2	2	4



Exercise

- Task 2

Naive Bayes

Naive Bayes

$$P(Y|\mathbf{X}) \propto P(Y) \prod_{i=1}^{n} P(X_i|Y)$$

Document is a sequence of terms $d = < t_1, \ldots, t_{|d|} >$

$$P(c|d) \propto P(c) \prod_{i=1}^{|d|} P(t_i|c)$$

Probability of a document being in class c

Prior probability of a document occurring in class c

Probability of a term given a class

Naive Bayes

- Document is a sequence of terms
  $d = < t_1, \ldots, t_{|d|} >$

$$P(c|d) \propto P(c) \prod_{i=1}^{|d|} P(t_i|c)$$

- Document is a **bag** of terms

$$P(c|d) \propto P(c) \prod_{t \in d} P(t|c)^{\boxed{n(t,d)}}$$

Number of times t occurs in d

For all terms in the document

Naive Bayes

- Prior probability
  - Relative frequency of class c in the training data

$$P(c|d) \propto \boxed{P(c)} \prod_{t \in d} P(t|c)^{n(t,d)}$$

$$\frac{N_c}{N}$$

N_c → Number of document in class c

N → Total number of documents

Naive Bayes

- Term probability
  - Multinomial distribution is a natural way to model distributions over frequency vectors
  - Terms occur zero or more times

$$P(c|d) \propto P(c) \prod_{t \in d} \boxed{P(t|c)}^{n(t,d)}$$

Relative frequency of the term in the class

$$\frac{n(t,c)}{\sum_{t'} n(t',c)}$$

Number of occurrences of t in training documents from class c

Sum of all term frequencies for class c

Naive Bayes

- Term probability
  - Multinomial distribution is a natural way to model distributions over frequency vectors
  - Terms occur zero or more times

$$P(c|d) \propto P(c) \prod_{t \in d} \boxed{P(t|c)}^{n(t,d)}$$

Relative frequency of the term in the class

What if this probability is zero?

$$\frac{n(t,c)}{\sum_{t'} n(t',c)}$$

Naive Bayes

- Term probability
  - Multinomial distribution is a natural way to model distributions over frequency vectors
  - Terms occur zero or more times

$$P(c|d) \propto P(c) \prod_{t \in d} \boxed{P(t|c)}^{n(t,d)}$$

Apply Laplace (or add-one) **smoothing**

$$\frac{n(t,c)+1}{\sum_{t'} n(t',c) + \boxed{|V|}}$$

Size of the vocabulary (number of distinct terms)

Example

	docID	terms						target
		chinese	beijing	shanghai	macao	tokyo	japan	(in China?)
training set	1	2	1					Yes
	2	2		1				Yes
	3	1			1			Yes
	4	1				1	1	No
test set	5	3				1	1	?

Probability of Yes class

P(Yes) * P(chinese|Yes)^3 * P(tokyo|Yes) * P(japan|Yes)

Probability of No class

P(No) * P(chinese|No)^3 * P(tokyo|No) * P(japan|No)

Example

class	N_c	n(t,c)						
		chinese	beijing	shanghai	macao	tokyo	japan	SUM
c=Yes	3	5	1	1	1			8
c=No	1	1				1	1	3

Probability of Yes class

P(Yes) * P(chinese|Yes)^3 * P(tokyo|Yes) * P(japan|Yes)

3/4 [(5+1)/(8+6)]^3=0.078 (0+1)/(8+6)=0.071 (0+1)/(8+6)=0.071 **=0.0003**

Probability of No class

P(No) * P(chinese|No)^3 * P(tokyo|No) * P(japan|No)

1/4 [(1+1)/(3+6)]^3=0.011 (1+1)/(3+6)=0.22 (1+1)/(3+6)=0.22 **=0.0001**

Exercise

- Task 3

Practical Issue

- Multiplying many small probabilities can result in numerical underflows
- In practice, log-probabilities are computed
  - Log is a monothonic transformation, does not change the outcome

$$P(c|d) \propto P(c) \prod_{t \in d} P(t|c)^{n(t,d)}$$

$$\Downarrow$$

$$\log P(c|d) \propto \log P(c) + \sum_{t} n(t,d) \log P(t|c)$$

Classification in Search Engines

- SPAM detection
- Sentiment analysis
  - Movie or product reviews as positive/negative
- Online advertising
- Vertical search

Text Clustering

Text Clustering

- As before, but using the notion of document similarity (Jaccard or cosine similarity)
- K-Means Clustering
- Hierarchical Agglomerative Clustering

K-Means Clustering

1. Select *K* points as initial centroids
2. **repeat**
   3. Form *K* clusters by assigning each point to its closest centroid
   4. Recompute the centroid of each cluster
5. **until** centroids do not change

K-Means Clustering

1. Select *K* points as initial centroids
2. **repeat**
   3. Form *K* clusters by assigning each point to its closest centroid
   4. Recompute the centroid of each cluster
5. **until** centroids do not change

Using Jaccard or cosine similarity

K-Means Clustering

1. Select *K* points as initial centroids
2. **repeat**
   3. Form *K* clusters by assigning each point to its closest centroid
   4. Recompute the centroid of each cluster
5. **until** centroids do not change

Taking the average term frequencies

Hierarchical Agglomerative Clustering

1. Compute the proximity matrix
2. **repeat**
   3. Merge the closest two clusters
   4. Update the proximity matrix
5. **until** only one cluster remains

Hierarchical Agglomerative Clustering

1. Compute the proximity matrix
2. **repeat**
   3. Merge the closest two clusters
   4. Update the proximity matrix
5. **until** only one cluster remains

Using Jaccard or cosine similarity

Hierarchical Agglomerative Clustering

1. Compute the proximity matrix
2. **repeat**
   3. $\boxed{\text{Merge the closest two clusters}}$
   4. Update the proximity matrix
5. **until** only one cluster remains

Taking the sum of term frequencies

Exercise

- Tasks 4 and 5