MONASH University

ETC3555

Statistical Machine Learning

Text mining

2 October 2018

Outline

Week	Topic
1	The learning problem
2	The learning problem
3	Linear models
4	Gradient descent
5	Neural Networks
6	Neural Networks
7	Deep Neural Networks
8	Support Vector Machines
9	Recommender systems
	Semester break
10	Text mining
11	Text mining and revision
12	Project presentation

Some references

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. **Introduction to Information Retrieval**, Cambridge University Press, 2008.
- Daniel Jurafsky and James H. Martin. **Speech and Language Processing**, Pearson Education, 2000.
- Charu C. Aggarwal and ChengXiang Zhai, **Mining Text Data**, Springer, 2012.

Outline

What is Text Mining?

"**Text mining**, also referred to as text data mining, roughly equivalent to text analytics, is the process of **deriving high-quality information from text**." (Wikipedia, 2018)

Text mining examples

- Sentiment analysis
- Document summarization
- Text Clustering
- Text Categorization
- Movie recommendation
- Restaurant/hotel recommendation
- News recommendation
- Text analytics in healthcare

How to perform text mining?

Text (data) mining = Text + Data Mining

- Text
    - Emails
    - Scientific literature
    - Tweets
    - News articles
- Data mining
    - Information retrieval (filter information)
    - Natural language processing (Tokenization, Part-of-speech tagging, etc)
    - Machine learning (knowledge discovery, predictive analytics, etc)

Some challenges in text mining

- Text data is not well-organized
  - Unstructured or semi-structured
- Natural language text contains ambiguities on many levels (syntactic, lexical etc)
  - The professor said on Monday he would give an exam.
  - He reached the bank
- Expensive to produce large-scale annotated training examples for learning

Outline

How to represent a document?

Monash University (/ˈmɒnæʃ/) is a public research university based in Melbourne, Australia. Founded in 1958, it is the second oldest university in the State of Victoria. The university has a number of campuses, four of which are in Victoria (Clayton, Caulfield, Peninsula, and Parkville), and one in Malaysia. Monash also has a research and teaching centre in Prato, Italy, a graduate research school in Mumbai, India and a graduate school in Suzhou, China. Monash University courses are also delivered at other locations, including South Africa.

Monash is home to major research facilities, including the Monash Law School, the Australian Synchrotron, the Monash Science Technology Research and Innovation Precinct (STRIP), the Australian Stem Cell Centre, 100 research centres[6] and 17 co-operative research centres. In 2016, its total revenue was over $2.2 billion dollars (AUD), with external research income around $282 million.[7] In 2016, Monash enrolled over 50,000 undergraduate and over 22,000 graduate students.[5] It has more applicants than any other university in the state of Victoria.[8]

Monash is a member of Australia's Group of Eight, a coalition of Australia's eight leading research Universities, a member of the ASAIHL, and is the only Australian member of the M8 Alliance of Academic Health Centers, Universities and National Academies. Monash is one of two Australian universities to be ranked in the École des Mines de Paris (Mines ParisTech) ranking on the basis of the number of alumni listed among CEOs in the 500 largest worldwide companies.[9]

- Represent by a string?
- Represent by a list of sentences?

How to represent a document?

Monash University (/ˈmɒnæʃ/) is a public research university based in Melbourne, Australia. Founded in 1958, it is the second oldest university in the State of Victoria. The university has a number of campuses, four of which are in Victoria (Clayton, Caulfield, Peninsula, and Parkville), and one in Malaysia. Monash also has a research and teaching centre in Prato, Italy, a graduate research school in Mumbai, India and a graduate school in Suzhou, China. Monash University courses are also delivered at other locations, including South Africa.

Monash is home to major research facilities, including the Monash Law School, the Australian Synchrotron, the Monash Science Technology Research and Innovation Precinct (STRIP), the Australian Stem Cell Centre, 100 research centres[6] and 17 co-operative research centres. In 2016, its total revenue was over $2.2 billion dollars (AUD), with external research income around $282 million.[7] In 2016, Monash enrolled over 50,000 undergraduate and over 22,000 graduate students.[5] It has more applicants than any other university in the state of Victoria.[8]

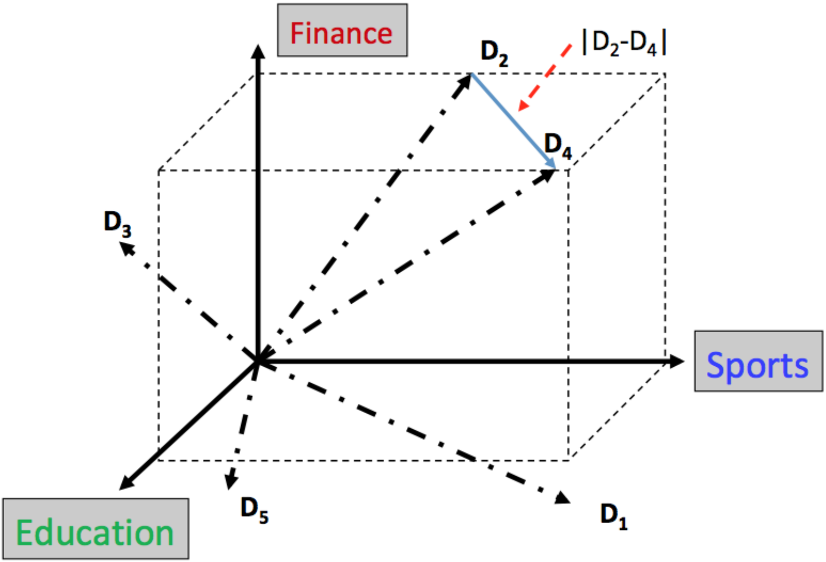
Monash is a member of Australia's Group of Eight, a coalition of Australia's eight leading research Universities, a member of the ASAIHL, and is the only Australian member of the M8 Alliance of Academic Health Centers, Universities and National Academies. Monash is one of two Australian universities to be ranked in the École des Mines de Paris (Mines ParisTech) ranking on the basis of the number of alumni listed among CEOs in the 500 largest worldwide companies.[9]

- String $\rightarrow$ No semantic meaning
- Sentence $\rightarrow$ A sentence is just another short document

Vector space model

- Represent documents by concept vectors.
- Each concept defines one dimension (multiple concepts $\rightarrow$ high-dimensional space).
- Each element of the vector corresponds to the concept weight, i.e. $D = (x_1, x_2, \ldots, x_k)^T$ where x_k is the importance of concept k in the document D.
- The relationship among documents is given by the distance between the concept vectors.

Vector space model



Vector space model

- How to define/select the concepts?
- Weights indicate how well the concept characterizes the document. How to assign weights?
- How to define the distance metric?

What is a good concept?

- Orthogonal basis vectors $\rightarrow$ Non-overlapping + No ambiguity
- Weights can be assigned automatically and accurately
- Some solutions
  - Terms or N-grams, a.k.a., Bag-of-Words
  - Topics

Bag-of-Words representation

- D_1: “Text mining is to identify useful information.”
- D_2: “Useful information is mined from text.”
- D_3: “Apple is delicious.”

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
Doc1	1	1	1	1	0	1	1	1	0	0	0
Doc2	1	1	0	0	1	1	1	0	1	0	0
Doc3	0	0	0	0	0	1	0	0	0	1	1

$\rightarrow$ Term as the basis for vector space

Tokenization

Break a stream of text into meaningful units/tokens (words, phrases, symbols, etc)

D: It’s not straightforward to perform so-called “tokenization”.

- “It’s”, “not”, “straightforward”, “to”, “perform”, “so-called”, “tokenization”
- “Ìt”, “ ’ ”, “s”, “not”, “straightforward”, “to”, “perform”, “so”, “-”, “called”, “‘ ” ”, “tokenization”, “.”, “ ” ”

Definition depends on language, corpus, or even context.

Bag-of-Words representation

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
Doc1	1	1	1	1	0	1	1	1	0	0	0
Doc2	1	1	0	0	1	1	1	0	1	0	0
Doc3	0	0	0	0	0	1	0	0	0	1	1

- It has the advantage of simplicity
- It assumes the words are independent from each other
- Grammar and order are missing

Bag-of-Words with N-grams

- N-grams: a contiguous sequence of N tokens from a given piece of text
- “Text mining is to identify useful information.”
  - text_mining, mining_is, is_to, to_identify, identify_useful, useful_information, information_.
- It captures local dependency and order
- Bias and variance tradeoff

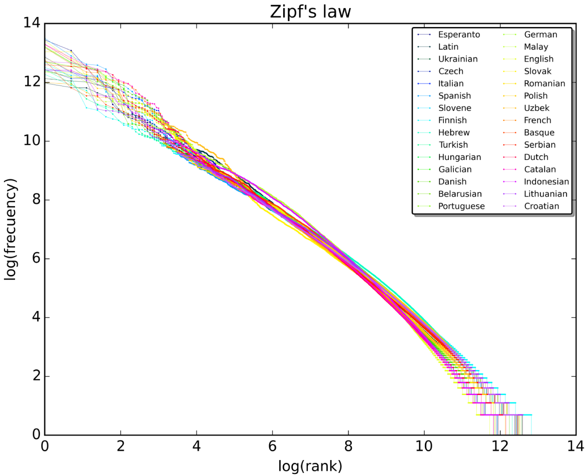
Should we represent the whole document?

- Represent a document with <u>all</u> the occurring words
- Pros
  - Preserve all information in the text
  - Fully automatic
- Cons
  - gap in the vocabulary: cars vs. car, talk vs. talking
  - Large storage

A statistical property of language

- **Zipf's law**: Frequency of any word is inversely proportional to its rank in the frequency table. In other words, the frequency f_i of the ith most common term is proportional to $\frac{1}{i}$, i.e. $f_i \propto \frac{1}{i}$
- The intuition is that frequency decreases very rapidly with rank

A statistical property of language

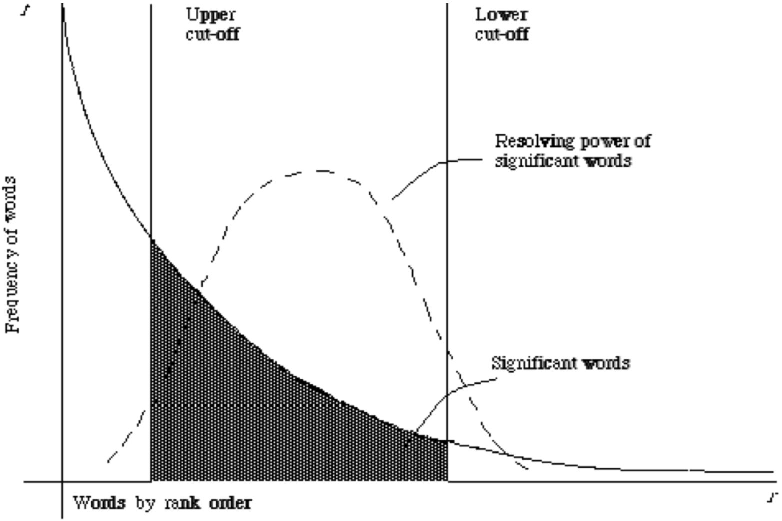


Rank versus frequency for the first 10 million words in 30 Wikipedias (dumps from October 2015) in a log-log scale.

A statistical property of language

- Head words take large portion of occurrences, but they are semantically meaningless, e.g. the, a, an, we, do, and to.
- Tail words take major portion of vocabulary, but they rarely occur in documents, e.g. dextrosinistral
- We should keep the middle words in our vocabulary since they are more representative

Non-informative and rare words



Normalization

- **Normalization**: convert different forms of a word to a normalized form in the vocabulary.
  - Example: U.S.A. $\rightarrow$ USA, St. Louis $\rightarrow$ Saint Louis
  - Rule-based (e.g. delete periods and hyphens, all in lower case) and Dictionary-based methods
- **Stemming**: reduce words to their root form
  - ladies $\rightarrow$ lady, referring $\rightarrow$ refer, forgotten $\rightarrow$ forget
  - Various stemming algorithms (Porter, Krovetz, etc)
  - There is a risk to lose precise meaning of the word.
  - Example: lay $\rightarrow$ lie (a false statement? or be in a horizontal position?)

Normalization

- **Remove stopwords** to reduce vocabulary size
  - Useless words for document analysis (not all words are informative)
  - No universal definition
  - Might break the original meaning/structure of text
  - Example: this is not a good option $\rightarrow$ option
  - Example: To be or not to be $\rightarrow$ null

Constructing a VSM representation

D_1: 'Text mining is to identify useful information.'

1. Tokenization
   - D_1: Text, mining, is, to, identify, useful, information, .
2. Stemming/normalization
   - D_1: text, mine, is, to, identify, use, inform, .
3. N-gram construction
   - D_1: text_mine, mine_is, is_to, to_identify, identify_use, use_inform, inform_.
4. Stopwords/vocabulary filtering
   - D_1: text_mine, to_identify, identify_use, use_inform

How to assign weights?

- Term Frequency (TF)
- Inverse Document Frequency (IDF)

Term frequency (TF)

- A term is more important if it occurs more frequently in a document
- Let $c(t, D)$ be the frequency count of term t in document $D \rightarrow \mathrm{TF}(t, D) = c(t, D)$
- Which documents are more similar to each other:
  - D_1: 'good', 10, D_2: 'good, 2, D_3: 'good', 3

Term frequency (TF)

- A term is more important if it occurs more frequently in a document
- Let $c(t, D)$ be the frequency count of term t in document $D \rightarrow \text{TF}(t, D) = c(t, D)$
- Which documents are more similar to each other:
  - D_1: 'good', 10, D_2: 'good, 2, D_3: 'good', 3
- Document length variation
- Twenty occurrences of a term in a document do not carry twenty times the significance of a single occurrence

$\rightarrow$ Term frequency normalization

Sub-linear TF normalization

$$\mathrm{TF}(t, D) = \begin{cases} 1 + log(c(t, D)), & c(t, D) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

- Penalize long document
- Various other TF normalization methods

Inverse document frequency (IDF)

- A term is more discriminative if it occurs only in fewer documents $\rightarrow$ Assign higher weights to the rare terms
- A corpus-specific property (independent of a single document)
- The IDF of a term t is computed as

$$\mathrm{IDF}(t) = 1 + \log\left(\frac{N}{df(t)}\right)$$

where

- N is the total number of documents in the collection/corpus,
- $\mathrm{df}(t)$ is the number of documents in the collection that contain the term t.

Document frequency vs collection frequency

$$\text{cf}(t) = \sum_{D} TF(t, D)$$

Word	cf	df
try	10422	8760
insurance	10440	3997

Cannot recognize words frequently occurring in a subset of documents

TF-IDF weighting

$$w(t, D) = \mathrm{TF}(t, D) \times \mathrm{IDF}(t)$$

- Combining TF and IDF
- Common in document $\rightarrow$ high TF $\rightarrow$ high weight
- Rare in collection $\rightarrow$ high IDF $\rightarrow$ high weight

TF-IDF assigns to term t a weight in document D that is

- highest when t occurs many times within a small number of documents;
- lower when the term occurs fewer times in a document, or occurs in many documents;
- lowest when the term occurs in virtually all documents.

Exercise

	D_1	D_2	D_3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

	df	idf
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

Which similarity metric?

Euclidean distance?

$$\text{dist}(D_1, D_2) = \sqrt{\sum_{t \in \text{Voc.}} \left(\underbrace{\text{TF}(t, D_1)\text{IDF}(t)}_{w(t,D_1)} - \underbrace{\text{TF}(t, D_2)\text{IDF}(t)}_{w(t,D_2)} \right)^2}$$

Which similarity metric?

Euclidean distance?

$$\mathrm{dist}(D_1, D_2) = \sqrt{\sum_{t \in \mathrm{Voc.}} \left(\underbrace{\mathrm{TF}(t, D_1)\mathrm{IDF}(t)}_{w(t,D_1)} - \underbrace{\mathrm{TF}(t, D_2)\mathrm{IDF}(t)}_{w(t,D_2)} \right)^2}$$

- Two documents with very similar content can have a significant vector difference simply because one is much longer than the other.
- The relative distributions of terms may be identical in the two documents, but the absolute term frequencies of one may be far larger.

Which similarity metric?

To compensate for the effect of document length, the standard way of quantifying the similarity between two documents D_1 and D_2 is to compute

$$\mathrm{sim}(D_1, D_2) = \frac{w_1^T w_2}{|w_1||w_2|} = \tilde{w}_1^T \tilde{w}_2,$$

where $w_j = w(D_j) = (w(t_1, D_j), w(t_2, D_j), \ldots, w(t_k, D_j))^T$ and $j = 1, 2$.

Which similarity metric?

To compensate for the effect of document length, the standard way of quantifying the similarity between two documents D_1 and D_2 is to compute

$$\text{sim}(D_1, D_2) = \frac{w_1^T w_2}{|w_1||w_2|} = \tilde{w}_1^T \tilde{w}_2,$$

where $w_j = w(D_j) = (w(t_1, D_j), w(t_2, D_j), \ldots, w(t_k, D_j))^T$ and $j = 1, 2$.

- Documents are normalized by length
- This is also called the *cosine similarity*

Exercise

- What is the range of values that the cosine similarity can take?

Exercise

- What is the range of values that the cosine similarity can take?
  - In general, the cosine has a range of $[-1, 1]$, but since all word counts are non-negative, the range is $[0, 1]$

Exercise

- What is the range of values that the cosine similarity can take?
  - In general, the cosine has a range of $[-1, 1]$, but since all word counts are non-negative, the range is $[0, 1]$
- D_1 : "Cat, dog, dog", $w(D_1) = (1, 2, 0)^T$ and D_2 : "$cat, dog, mouse, mouse$", $w(D_2) = (1, 1, 2)^T$. Cosine similarity between D_1 and D_2? Between D_1 and D_1?

Exercise

- What is the range of values that the cosine similarity can take?
  - In general, the cosine has a range of $[-1, 1]$, but since all word counts are non-negative, the range is $[0, 1]$
- D_1 : “Cat, dog, dog'', $w(D_1) = (1, 2, 0)^T$ and D_2 : “$cat, dog, mouse, mouse''$, $w(D_2) = (1, 1, 2)^T$. Cosine similarity between D_1 and D_2? Between D_1 and D_1?
  - 0.55 and 1.

Exercise

- What is the range of values that the cosine similarity can take?
  - In general, the cosine has a range of $[-1, 1]$, but since all word counts are non-negative, the range is $[0, 1]$
- D_1 : "Cat, dog, dog", $w(D_1) = (1, 2, 0)^T$ and D_2 : "$cat, dog, mouse, mouse$", $w(D_2) = (1, 1, 2)^T$. Cosine similarity between D_1 and D_2? Between D_1 and D_1?
  - 0.55 and 1.
- $w(D_1) = (1, 3)^T$, $w(D_2) = (10, 30)^T$ and $w(D_3) = (3, 1)^T$. Cosine similarity between D_1 and D_2, and D_2 and D_3?

Exercise

- What is the range of values that the cosine similarity can take?
    - In general, the cosine has a range of $[-1, 1]$, but since all word counts are non-negative, the range is $[0, 1]$
- D_1 : "Cat, dog, dog", $w(D_1) = (1, 2, 0)^T$ and D_2 : "$cat, dog, mouse, mouse$", $w(D_2) = (1, 1, 2)^T$. Cosine similarity between D_1 and D_2? Between D_1 and D_1?
    - 0.55 and 1.
- $w(D_1) = (1, 3)^T$, $w(D_2) = (10, 30)^T$ and $w(D_3) = (3, 1)^T$. Cosine similarity between D_1 and D_2, and D_2 and D_3?
    - 1 and 0.6.

Exercise

x and y are two k-dimensional unit vectors. What is the relationship between the cosine similarity and the euclidean distance? (Hint: can you compute cosine similarity from Euclidean distance, and vice versus?)

Exercise

x and y are two k-dimensional unit vectors. What is the relationship between the cosine similarity and the euclidean distance? (Hint: can you compute cosine similarity from Euclidean distance, and vice versus?)

$$\sqrt{\sum_{j=1}^{k}(x_j - y_j)^2} = \sqrt{\sum_{j=1}^{k} x_j^2 + \sum_{j=1}^{k} y_j^2 - 2\sum_{j=1}^{k} x_j y_j}$$

$$= \sqrt{2 - 2\sum_{j=1}^{k} x_j y_j} = \sqrt{2 - \frac{2\sum_{j=1}^{k} x_j y_j}{\sqrt{\sum_{j=1}^{k} x_j^2}\sqrt{\sum_{j=1}^{k} y_j^2}}}$$

$$= \sqrt{2 - 2 \times \cos(x, y)}$$

The vector space model

- Advantages
  - Simple (intuitive, easy to implement, etc)
  - Empirically effective
  - Well-studied
- Disadvantages
  - Assume term independence
  - Arbitrary term weighting
  - Arbitrary similarity measure
  - Multiple parameter tuning