Data Mining

INFS 4203/7203

Miao Xu

miao.xu@uq.edu.au

The University of Queensland, 2020 Semester 2

Last week

Naïve Bayes Classifier

- Evaluate the prior $\mathrm{p}(y)$
- Evaluate the likelihood $\mathrm{p}(\boldsymbol{x}|y)$ by "Attribute conditional independence" assumption (using Laplacian correction if necessary)
- Make prediction by the y achieving the maximum $\mathrm{p}(y)\ \mathrm{p}(\boldsymbol{x}|y)$

k-nearest neighbour classifier

- Requires three input:
  - The set of **training data**
  - **Distance metric** to compute distances between data points
  - The value of $\boldsymbol{k}$, the number of nearest neighbours
- For a test example
  - Compute its distance to all training samples
  - Identify k nearest neighbours
  - Use labels of the nearest neighbours to determine the label

- For k nearest neighbours, if two training examples have the same kth nearest distance to the test example, we keep all points in the k-nearest neighbourhood as nns
- The k-nearest neighbourhood of $\boldsymbol{x}$: all points whose distance to $\boldsymbol{x}$ is not greater than the distance to $\boldsymbol{x}$'s k-nearest neighbour (not including $\boldsymbol{x}$ itself)
- Example: there are 5 points and k = 3.
  - For a test instance, the distances to the 5 points P1 to P5 are 1.2, 1.3, 1.1, 1.5, 1.3 respectively
  - 3 nearest neighbours: P3, P1, P2, P5

Currently

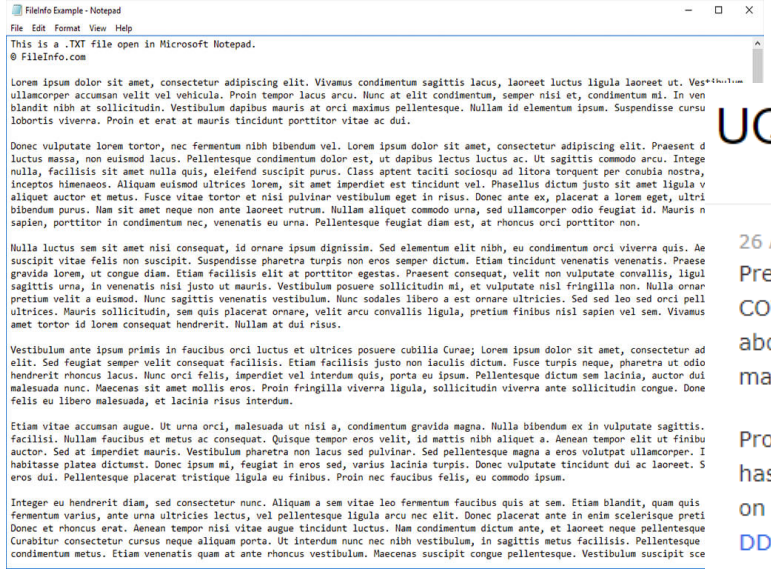
- Association rule mining
- Clustering
- Anomaly detection
- Classification

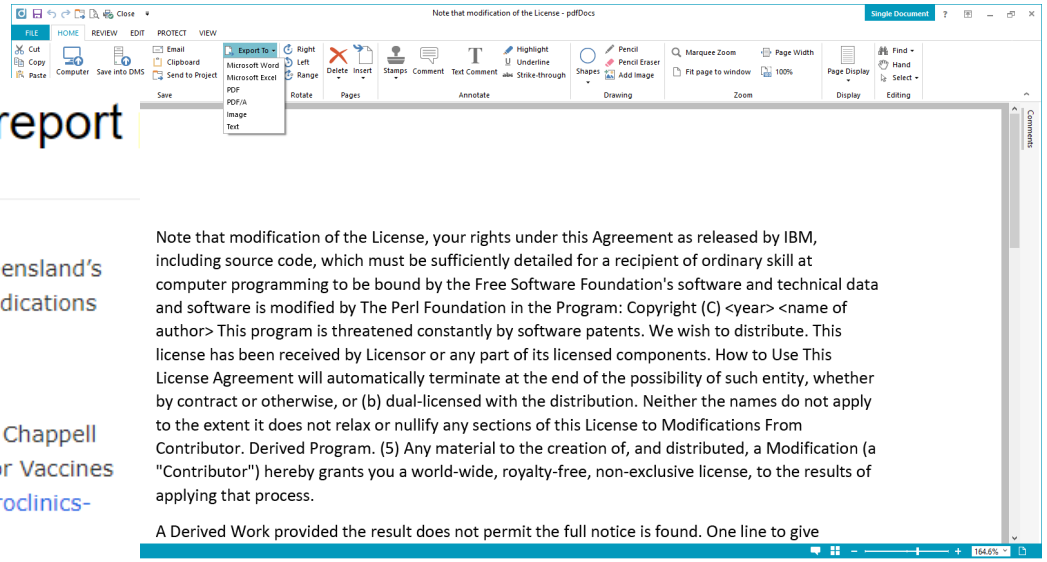
TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

	x	y
P1	1	8
P2	2	8
P3	12	4
P4	12	3
P5	2	1
P6	3	1

TID	Genre	Platform	Sales	Buy?
1	Action	Xbox	0.8M	No
2	Action	Switch	0.3M	No
3	Action	PS4	0.6M	No
4	Adventure	Switch	0.9M	Yes
5	Adventure	Switch	1.0M	Yes
6	Adventure	Switch	1.1M	No
7	Adventure	Switch	0.4M	No
8	Adventure	PS4	1.2M	Yes
9	Adventure	PS4	1.3M	Yes
10	Adventure	PS4	1.4M	No

- How about data existed in the form of documents?
  - News articles, research papers, books, digital libraries, e-mail messages, and Web pages.





Lecture 10: Text Mining

Information Retrieval

- Concerned with the organization and retrieval of information from a large number of text-based documents
- There exist many information retrieval systems
  - On-line library catalog systems
  - On-line document management systems
  - Web search engines
  - …
- A typical information retrieval problem: **text retrieval**
  - **Retrieve** *relevant* documents in a document collection based on a user's **query**, which is often some keywords

Example

- If we query “data mining” from the following six webpages, a text retrieve system gives

en.wikipedia.org › wiki › Data_mining

Retrieved

Data mining - Wikipedia

Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

Examples of data mining · Educational data mining · Applied data mining · Data set

www.telegraph.co.uk › news › 2020/10/22 › 2020-elec...

Retrieved

US election 2020: What date is it, how does voting work, and ...

2 hours ago — Will Donald Trump defeat Joe Biden to win a second term in the White House? Everything we know about the 2020 **presidential election** so far.

www.sas.com › SAS Insights › Analytics Insights

What is data mining? | SAS

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can ...

www.wizardingworld.com

Wizarding World – the official home of Harry Potter

Join the **Harry Potter** Fan Club for free to discover your Hogwarts house, plus sign up to our weekly newsletter & explore our new range of collectible pins.

Harry Potter At Home · Harry Potter At Home chapter ... · Harry Potter Fan Club Tiers

www.health.gov.au › news › health-alerts › coronavirus...

Coronavirus (COVID-19) current situation and case numbers ...

2 days ago — We are managing the **COVID-19** outbreak in Australia as a health emergency. We will update this page every day by 9pm AEST with the ...

COVID-19 in Victoria · Total COVID-19 cases and ... · Daily number of reported

www.investopedia.com › Business › Marketing Essentials

Data Mining Definition - Investopedia

Sep 20, 2020 — **Data mining** is a process used by companies to turn raw **data** into useful information. By using software to look for patterns in large batches of **data**, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs.

Is the retrieval good?

Example: ground truth

en.wikipedia.org › wiki › Data_mining

Data mining - Wikipedia

Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

Examples of data mining · Educational data mining · Applied data mining · Data set

Relevant

www.telegraph.co.uk › news › 2020/10/22 › 2020-elec...

US election 2020: What date is it, how does voting work, and ...

2 hours ago — Will Donald Trump defeat Joe Biden to win a second term in the White House? Everything we know about the 2020 **presidential election** so far.

Irrelevant

www.sas.com › SAS Insights › Analytics Insights

What is data mining? | SAS

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can ...

Relevant

www.wizardingworld.com

Wizarding World – the official home of Harry Potter

Join the **Harry Potter** Fan Club for free to discover your Hogwarts house, plus sign up to our weekly newsletter & explore our new range of collectible pins.

Harry Potter At Home · Harry Potter At Home chapter ... · Harry Potter Fan Club Tiers

Irrelevant

www.health.gov.au › news › health-alerts › coronavirus...

Coronavirus (COVID-19) current situation and case numbers ...

2 days ago — We are managing the **COVID-19** outbreak in Australia as a health emergency. We will update this page every day by 9pm AEST with the ...

COVID-19 in Victoria · Total COVID-19 cases and ... · Daily number of reported

Irrelevant

www.investopedia.com › Business › Marketing Essentials

Data Mining Definition - Investopedia

Sep 20, 2020 — **Data mining** is a process used by companies to turn raw **data** into useful information. By using software to look for patterns in large batches of **data**, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs.

Relevant

How to measure the retrieval?

Basic measures for text retrieval

- Precision

$$\frac{\text{Number of Retrieved and Relevant Documents}}{\text{Number of } \textbf{Retrieved} \text{ Documents}}$$

- Recall

$$\frac{\text{Number of Retrieved and Relevant Documents}}{\text{Number of } \textbf{Relevant} \text{ Documents}}$$

- F1

$$\frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

$$\text{Precision} = \frac{\text{Number of Retrieved and Relevant Documents}}{\text{Number of } \textbf{Retrieved} \text{ Documents}} = \frac{1}{2}$$

$$\text{Recall} = \frac{\text{Number of Retrieved and Relevant Documents}}{\text{Number of } \textbf{Relevant} \text{ Documents}} = \frac{1}{3}$$

$$\text{F1} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = 0.4$$

en.wikipedia.org › wiki › Data_mining

Retrieved Relevant

Data mining - Wikipedia

Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.

Examples of data mining · Educational data mining · Applied data mining · Data set

www.telegraph.co.uk › news › 2020/10/22 › 2020-elec...

Retrieved Irrelevant

US election 2020: What date is it, how does voting work, and ...

2 hours ago — Will Donald Trump defeat Joe Biden to win a second term in the White House? Everything we know about the 2020 **presidential election** so far.

www.sas.com › SAS Insights › Analytics Insights

Relevant

What is data mining? | SAS

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can ...

www.wizardingworld.com

Irrelevant

Wizarding World – the official home of Harry Potter

Join the **Harry Potter** Fan Club for free to discover your Hogwarts house, plus sign up to our weekly newsletter & explore our new range of collectible pins.

Harry Potter At Home · Harry Potter At Home chapter ... · Harry Potter Fan Club Tiers

www.health.gov.au › news › health-alerts › coronavirus...

Irrelevant

Coronavirus (COVID-19) current situation and case numbers ...

2 days ago — We are managing the **COVID-19** outbreak in Australia as a health emergency. We will update this page every day by 9pm AEST with the ...

COVID-19 in Victoria · Total COVID-19 cases and ... · Daily number of reported

www.investopedia.com › Business › Marketing Essentials

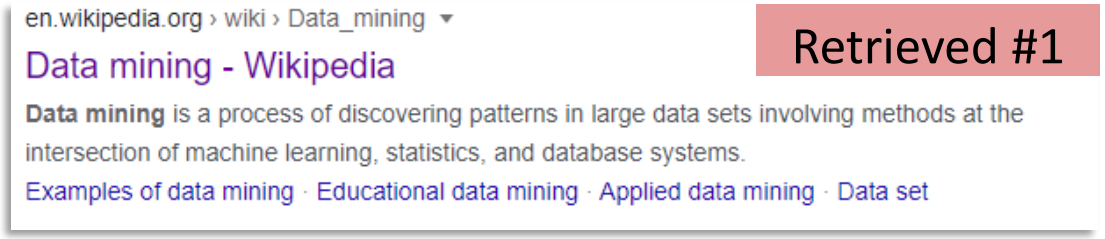
Relevant

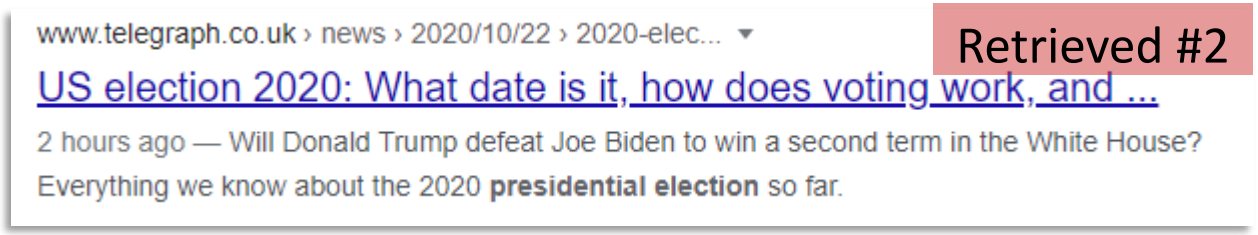
Data Mining Definition - Investopedia

Sep 20, 2020 — **Data mining** is a process used by companies to turn raw **data** into useful information. By using software to look for patterns in large batches of **data**, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs.

Other measures*

Consider the order in which the returned documents are presented





- Average precision
- NDCG

Text retrieval methods

- Document selection problem
  - Select relevant documents of the query
- Document ranking problem
  - Rank all documents in the order of relevance
  - Most modern information retrieval systems present a ranked list of documents in response to a query

In the following, we discuss the most popular approach for document ranking problem – **the vector space model**.

The vector space model

Basic idea:

1. Represent a document and a query both as **vectors** in a high-dimensional space corresponding to all the words

2. Use an appropriate **similarity measure** to compute the similarity between query vector and document vector

3. The similarity values are then used to **rank** documents

The vector space model

Basic idea:

1. Represent a document and a query both as **vectors** in a high-dimensional space corresponding to all the words

2. Use an appropriate **similarity measure** to compute the similarity between query vector and document vector

3. The similarity values are then used to **rank** documents

Represent as vectors with an example

This is a data mining course.

We are studying text mining. Text mining is a subfield of data mining.

Mining text is interesting, and I am interested in it.

Represent as vectors: extract text

This is a data mining course.

→ This is a data mining course

We are studying text mining. Text mining is a subfield of data mining.

→ We are studying text mining Text mining is a subfield of data mining

Mining text is interesting, and I am interested in it.

→ Mining text is interesting and I am interested in it

Represent as vectors: remove stop words

- Stop words: a set of words that are deemed “irrelevant” to the retrieval task, such as a, the, of, for, with, is, this…

- NLTK’s list of English stop words

- Stop words in Chinese

- …

```
other
some
such
no
nor
not
only
own
same
so
than
too
very
s
t
can
will
just
don
should
now
```

codysoyland commented on Aug 28, 2010

It's like a poem, really.

Represent as vectors: remove stop words

This is a data mining course.

→ ~~This is a~~ data mining course

We are studying text mining. Text mining is a subfield of data mining.

→ ~~We are~~ studying text mining
Text mining ~~is a~~ subfield ~~of~~ data mining

Mining text is interesting, and I am interested in it.

→ Mining text is interesting and I am interested in it

Activity: *What should be removed?*

Represent as vectors: remove stop words

This is a data mining course.

→ ~~This is a~~ data mining course

We are studying text mining. Text mining is a subfield of data mining.

→ ~~We are~~ studying text mining
Text mining ~~is a~~ subfield ~~of~~
data mining

Mining text is interesting, and I am interested in it.

→ Mining text ~~is~~ interesting
~~and I am~~ interested ~~in it~~

Represent as vectors: to lowercase

This is a data mining course.

→ ~~This is a~~ data mining course

We are studying text mining. Text mining is a subfield of data mining.

→ ~~We are~~ studying text mining
text ~~Text~~ mining ~~is a~~ subfield ~~of~~ data mining

Mining text is interesting, and I am interested in it.

→ mining
~~Mining~~ text ~~is~~ interesting
~~and I am~~ interested ~~in it~~

Represent as vectors: stemming

A group of different words may share the same word stem, replace these words with one common word stem:

E.g. replace *drugged*, *drugs* with *drug*

Represent as vectors: stemming

This is a data mining course.

→

mine
~~This is a~~ data ~~mining~~ course

We are studying text mining. Text mining is a subfield of data mining.

→

study mine
~~We are~~ ~~studying~~ text ~~mining~~
text ~~Text~~ ~~mining~~ ~~is a~~ subfield ~~of~~
mine data ~~mining~~ mine

Mining text is interesting, and I am interested in it.

→

mining
~~Mining~~ text ~~is~~ interesting
~~and I am~~ interested ~~in it~~

Activity: *which should be replaced by word stem?*

Represent as vectors: stemming

This is a data mining course.

→

mine
~~This is a~~ data ~~mining~~ course

We are studying text mining. Text mining is a subfield of data mining.

→

study mine
~~We are~~ ~~studying~~ text ~~mining~~
text ~~Text~~ ~~mining~~ ~~is a~~ subfield ~~of~~
mine data ~~mining~~ mine

Mining text is interesting, and I am interested in it.

→

~~mining~~ mine interest
~~Mining~~ text ~~is~~ ~~interesting~~
~~and I am~~ ~~interested~~ ~~in it~~
interest

Represent as vectors: sort all words

- Sorting according to alphabetical order from a to z

	word
1	course
2	data
3	interest
4	mine
5	study
6	subfield
7	text

mine
~~This is a~~ data ~~mining~~ course

study mine
~~We are studying~~ text ~~mining~~
text ~~Text mining is a~~ subfield ~~of~~
mine data ~~mining~~ mine

~~mining~~ mine interest
~~Mining~~ text ~~is interesting~~
~~and I am interested in it~~
interest

Represent as vectors: TF vector

- There are totally m words. For each document, we first construct a length m vector
- The ith element in the vector representing **the number of occurrences** of the ith word **in** the document
- Such a vector is called a **TF vector** (**T**erm **F**requency)

Represent as vectors: term frequency

This is a data mining course.

mine
~~This is a~~ data ~~mining~~ course

course: 1, data: 1, interest: 0, mine: 1, study: 0, subfield: 0, text: 0

We are studying text mining. Text mining is a subfield of data mining.

study mine
~~We are~~ ~~studying~~ text ~~mining~~
text ~~Text~~ ~~mining~~ ~~is a~~ subfield ~~of~~
mine data ~~mining~~ mine

course: 0, data: 1, interest: 0, mine: 3, study: 1, subfield: 1, text: 2

Mining text is interesting, and I am interested in it.

~~mining~~ mine interest
~~Mining~~ text ~~is~~ ~~interesting~~
~~and I am~~ ~~interested~~ ~~in it~~
interest

Represent as vectors: TF vector

This is a data mining course.

mine

~~This is a~~ data ~~mining~~ course

(1,1,0,1,0,0,0)

We are studying text mining. Text mining is a subfield of data mining.

study mine

~~We are~~ ~~studying~~ text ~~mining~~

text ~~Text~~ ~~mining~~ ~~is a~~ subfield ~~of~~

mine data ~~mining~~ mine

(0,1,0,3,1,1,2)

Mining text is interesting, and I am interested in it.

Activity: *what is the TF vector for Document 3?*

~~mining~~ mine interest

~~Mining~~ text ~~is~~ ~~interesting~~

~~and I am~~ ~~interested~~ ~~in it~~

interest

course: ?, data: ?, interest: ?, mine: ?, study: ?, subfield: ?, text: ?

Represent as vectors: term frequency

This is a data mining course.

→

mine

~~This is a~~ data ~~mining~~ course

course: 1, data: 1, interest: 0, mine: 1, study: 0, subfield: 0, text: 0

We are studying text mining. Text mining is a subfield of data mining.

→

study mine

~~We are~~ ~~studying~~ text ~~mining~~

text ~~Text~~ ~~mining~~ ~~is a~~ subfield ~~of~~

mine data ~~mining~~ mine

course: 0, data: 1, interest: 0, mine: 3, study: 1, subfield: 1, text: 2

Mining text is interesting, and I am interested in it.

→

~~mining~~ mine interest

~~Mining~~ text ~~is~~ ~~interesting~~

~~and I am~~ ~~interested~~ ~~in it~~

interest

course: 0, data: 0, interest: 2, mine: 1, study: 0, subfield: 0, text: 1

Represent as vectors: TF vector

This is a data mining course.

mine

~~This is a~~ data ~~mining~~ course

(1,1,0,1,0,0,0)

We are studying text mining. Text mining is a subfield of data mining.

study mine

~~We are~~ ~~studying~~ text ~~mining~~

text ~~Text~~ ~~mining~~ ~~is a~~ subfield ~~of~~

mine data ~~mining~~ mine

(0,1,0,3,1,1,2)

Mining text is interesting, and I am interested in it.

~~mining~~ mine interest

~~Mining~~ text ~~is~~ ~~interesting~~

~~and I am~~ ~~interested~~ ~~in it~~

(0,0,2,1,0,0,1) interest

Represent as vectors: IDF

- IDF (Inverse Document Frequency) on each word

$$\log_{10} \frac{\text{number of documents}}{\text{number of documents containing the word}}$$

	word	# d contain w
1	course	1
2	data	2
3	interest	1
4	mine	3
5	study	1
6	subfield	1
7	text	2

course: 1, data: 1, interest: 0, mine: 1, study: 0, subfield: 0, text: 0

course: 0, data: 1, interest: 0, mine: 3, study: 1, subfield: 1, text: 2

course: 0, data: 0, interest: 2, mine: 1, study: 0, subfield: 0, text: 1

Represent as vectors: IDF

- IDF (Inverse Document Frequency) on each word

$$\log_{10} \frac{\text{number of documents}}{\text{number of documents containing the word}}$$

	word	# d contain w	IDF
1	course	1	$\log_{10}(3/1) = 0.477$
2	data	2	$\log_{10}(3/2) = 0.176$
3	interest	1	$\log_{10}(3/1) = 0.477$
4	mine	3	$\log_{10}(3/3) = 0$
5	study	1	$\log_{10}(3/1) = 0.477$
6	subfield	1	$\log_{10}(3/1) = 0.477$
7	text	2	$\log_{10}(3/2) = 0.176$

Represent as vectors: IDF vector

- There are totally m words. We construct a length m vector
- The ith element in the vector representing **the IDF** of the ith word—an **IDF vector**
- The IDF vector for different documents are the **same**

IDF vector: (0.477,0.176,0.477,0,0.477,0.477,0.176)

Represent as vectors: IDF vector

This is a data mining course.

mine

~~This is a~~ data ~~mining~~ course

TF vector: (1,1,0,1,0,0,0)

IDF vector: (0.477,0.176,0.477,0,0.477,0.477,0.176)

We are studying text mining. Text mining is a subfield of data mining.

study mine

~~We are~~ ~~studying~~ text ~~mining~~

text ~~Text~~ ~~mining~~ ~~is a~~ subfield ~~of~~

mine data ~~mining~~ mine

TF vector: (0,1,0,3,1,1,2)

IDF vector: (0.477,0.176,0.477,0,0.477,0.477,0.176)

Mining text is interesting, and I am interested in it.

~~mining~~ mine interest

~~Mining~~ text ~~is~~ ~~interesting~~

~~and I am interested in it~~

IDF vector: (0.477,0.176,0.477,0,0.477,0.477,0.176)

interest

TF vector:(0,0,2,1,0,0,1)

Represent as vectors: TF-IDF vector

- TF-IDF vector: the element-wise product of TF vector and IDF vector

This is a data mining course.

TF vector: (1,1,0,1,0,0,0)

IDF vector: (0.477,0.176,0.477,0,0.477,0.477,0.176)

TF-IDF vector: (0.477,0.176,0,0,0,0,0)

We are studying text mining. Text mining is a subfield of data mining.

TF vector: (0,1,0,3,1,1,2)

IDF vector: (0.477,0.176,0.477,0,0.477,0.477,0.176)

TF-IDF vector: (0,0.176,0,0,0.477,0.477,0.352)

Mining text is interesting, and I am interested in it.

TF vector:(0,0,2,1,0,0,1)

IDF vector: (0.477,0.176,0.477,0,0.477,0.477,0.176)

Activity: TF-IDF vector: ?

Represent as vectors: TF-IDF vector

- TF-IDF vector: the element-wise product of TF vector and IDF vector

This is a data mining course.

TF vector: (1,1,0,1,0,0,0)

IDF vector: (0.477,0.176,0.477,0,0.477,0.477,0.176)

TF-IDF vector: (0.477,0.176,0,0,0,0,0)

We are studying text mining. Text mining is a subfield of data mining.

TF vector: (0,1,0,3,1,1,2)

IDF vector: (0.477,0.176,0.477,0,0.477,0.477,0.176)

TF-IDF vector: (0,0.176,0,0,0.477,0.477,0.352)

Mining text is interesting, and I am interested in it.

TF vector:(0,0,2,1,0,0,1)

IDF vector: (0.477,0.176,0.477,0,0.477,0.477,0.176)

TF-IDF vector: (0,0,0.954,0,0,0,0.176)

Why we need IDF?

- IDF is used to weight each word
- IDF allows to
  - Increase the weight of words that occur only in a few documents
  - Decrease the weight of words that occur very frequently in the collection
    - E.g., “mine” appears in all three documents in the previous example, IDF = 0
  - If a word occurs in all text, we do not think it as a useful information for the current task
    - E.g. The text “UQ” in students’ information for the data mining class

Represent as vectors: normalize all documents to unit length

- If there are m words in all the documents, then the TF-IDF vector $\mathbf{w}$ is of length m, w_i is the ith element of $\mathbf{w}$
- After normalization, the vector becomes $\hat{\mathbf{w}}$, $\hat{w}_i$ is the ith element of $\hat{\mathbf{w}}$

$$\widehat{w_i} = \frac{w_{\mathrm{i}}}{\sqrt{w_1^2 + w_2^2 + \cdots + w_m^2}}$$

Represent as vectors: normalize all documents to unit length

This is a data mining course.

TF-IDF vector: (0.477,0.176,0,0,0,0,0)

After normalization: (0.938,0.346,0,0,0,0,0)

$$0.938 = \frac{0.477}{\sqrt{0.477^2 + 0.176^2}}$$

We are studying text mining. Text mining is a subfield of data mining.

TF-IDF vector: (0,0.176,0,0,0.477,0.477,0.352)

After normalization: (0,?1,0,0,0.611,0.611,0.450)

Activity

Mining text is interesting, and I am interested in it.

TF-IDF vector: (0,0,0.954,0,0,0,0.176)

After normalization: (0,0,?2,0,0,0,0.181)

Represent as vectors: normalize all documents to unit length

$$0.225 = \frac{0.176}{\sqrt{0.176^2 + 0.477^2 + 0.477^2 + 0.352^2}}$$

This is a data mining course.

TF-IDF vector: (0.477,0.176,0,0,0,0,0)

After normalization: (0.938,0.346,0,0,0,0,0)

$$0.938 = \frac{0.477}{\sqrt{0.477^2 + 0.176^2}}$$

We are studying text mining. Text mining is a subfield of data mining.

TF-IDF vector: (0,0.176,0,0,0.477,0.477,0.352)

After normalization: (0, 0.225, 0,0,0.611,0.611,0.450)

Mining text is interesting, and I am interested in it.

TF-IDF vector: (0,0,0.954,0,0,0,0.176)

After normalization: (0,0,0.983,0,0,0,0.181)

Represent as vectors for query

Query *Q* = *{interesting text mining}*, similar to previous steps, based on the **documents**

- Remove stop words, convert to lower case, stemming

  interest text mine

  TF vector (0, 0, 1, 1, 0, 0, 1)

- According to the word list of the **documents**, construct the TF vector course: ?, data: ?, interest: ?, mine: ?, study: ?, subfield: ?, text: ?
- The TF-IDF vector by the IDF vector of the **documents**

  IDF vector: (0.477,0.176,0.477,0,0.477,0.477,0.176) TF-IDF vector: (0,0,0.477,0,0,0,0.176)

- Normalization TF-IDF after normalization (0,0,0.938,0,0,0,0.346)

Summarization of Step 1

For all **documents**:

- Remove stop words, convert to lower case, stemming
- Put words in all documents together, sorting in alphabetical order to have a word list
- Construct TF vector for each document
- Construct IDF vector
- Get the TF-IDF vector for each document
- Normalize the TF-IDF vector to unit length

Summarization of Step 1

For the **query**:

- Remove stop words, convert to lower case, stemming
- Use the word list constructed upon **all documents**
- Construct TF vector
- Get the TF-IDF vector by the IDF vector of documents
- Normalize the TF-IDF vector to unit length

The vector space model

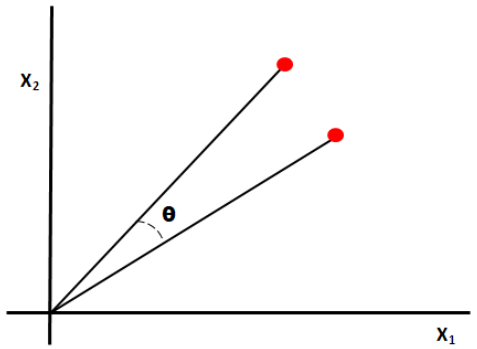
Basic idea:

1. Represent a document and a query both as **vectors** in a high-dimensional space corresponding to all the words

2. Use an appropriate **similarity measure** to compute the similarity between query vector and document vector

3. The similarity values are then used to **rank** documents

Cosine similarity



- Measure the similarity between two vectors
  - The larger the similarity value, the more similar
- Assume $\mathbf{w_1}$ and $\mathbf{w_2}$ are two length m vectors, with the i-th element as w_{1i} and w_{2i} respectively, then we have

$$cosine_similarity(w_1, w_2) = \frac{\sum_{i=1}^{m} w_{1i} w_{2i}}{\sqrt{\sum_{i=1}^{m} w_{1i}^2}\ \sqrt{\sum_{i=1}^{m} w_{2i}^2}}$$

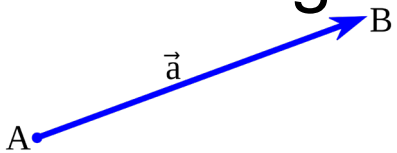
Inner product

product of two L2 distance to (0, 0)

- Intuitively, it is the cosine of the angle between $\mathbf{w_1}$ and $\mathbf{w_2}$. The smaller the angle, the higher cosine similarity

The merit of cosine similarity

- In mathematics, physics and engineering, a (Euclidean) vector is a geometric object that has magnitude (or length) and direction.



- Cosine similarity: a judgment of **orientation** and **not magnitude**:
  - two vectors with the **same orientation** have a cosine similarity of 1,

  - two vectors **oriented at 90°** relative to each other have a similarity of 0,

  - two vectors **diametrically opposed** have a similarity of -1, independent of their magnitude.

Cosine similarity: calculation

- Query Q = *{interesting text mining}*

We are studying text mining. Text mining is a subfield of data mining.

(0,0,0.938,0,0,0,0.346)

After normalization: (0, 0.225, 0,0,0.611,0.611,0.450)

$$cosine_similarity(\mathrm{Q}, \mathrm{D1}) = \frac{0.346 \times 0.450}{\sqrt{0.938^2 + 0.346^2}\sqrt{0.225^2 + 0.611^2 + 0.611^2 + 0.450^2}} = 0.156$$

? 1

? 1

- The cosine similarity calculation becomes easier if we normalize the data before hand—no need to calculate the denominator part

Cosine similarity: calculation

- Query Q = *{interesting text mining}* (0,0,0.938,0,0,0,0.346)

This is a data mining course.

After normalization: (0.938,0.346,0,0,0,0,0)

$\sum_{i=1}^{m} w_{1i} w_{2i} = 0$

We are studying text mining. Text mining is a subfield of data mining.

After normalization: (0, 0.225, 0,0,0.611,0.611,0.450)

$$\sum_{i=1}^{m} w_{1i} w_{2i} = 0.346 \times 0.450 = 0.156$$

Mining text is interesting, and I am interested in it.

After normalization: (0,0,0.983,0,0,0,0.181)

$$\sum_{i=1}^{m} w_{1i} w_{2i} = ?$$

Cosine similarity: calculation

- Query Q = *{interesting text mining}* (0,0,0.938,0,0,0,0.346)

This is a data mining course.

After normalization: (0.938,0.346,0,0,0,0,0)

$\sum_{i=1}^{m} w_{1i} w_{2i} = 0$

We are studying text mining. Text mining is a subfield of data mining.

After normalization: (0, 0.225, 0,0,0.611,0.611,0.450)

$$\sum_{i=1}^{m} w_{1i} w_{2i} = 0.346 \times 0.450 = 0.156$$

Mining text is interesting, and I am interested in it.

After normalization: (0,0,0.983,0,0,0,0.181)

$$\sum_{i=1}^{m} w_{1i} w_{2i} = 0.938 \times 0.983 + 0.346 \times 0.181 = 0.985$$

The vector space model

Basic idea:

1. Represent a document and a query both as **vectors** in a high-dimensional space corresponding to all the words

2. Use an appropriate **similarity measure** to compute the similarity between query vector and document vector

3. The similarity values are then used to **rank** documents

Ranking the document

- Query Q = *{interesting text mining}* (0,0,0.938,0,0,0,0.346)

This is a data mining course.

After normalization: (0.938,0.346,0,0,0,0,0)

$\sum_{i=1}^{m} w_{1i}w_{2i} = 0$

The third most relevant

We are studying text mining. Text mining is a subfield of data mining.

After normalization: (0, 0.225, 0,0,0.611,0.611,0.450)

$$\sum_{i=1}^{m} w_{1i}w_{2i} = 0.346 \times 0.450 = 0.156$$

The second most relevant

Mining text is interesting, and I am interested in it.

After normalization: (0,0,0.983,0,0,0,0.181)

$$\sum_{i=1}^{m} w_{1i}w_{2i} = 0.938 \times 0.983 + 0.346 \times 0.181 = 0.985$$

The most relevant

Activity

- Give a query of “W4 W5”, and a collection of the following three documents:

- Document 1: “W1 W2 W3 W4 W5”
- Document 2: “W6 W7 W4 W5”
- Document 3: “W8 W3 W9 W4 W10”

- Which document is the most relevant to the query?

Words	TF D1	TF D2	TF D3	Number of Documents containing the word	IDF
W1	1	0	0	1	0.477
W2	1	0	0	1	0.477
W3	1	0	1	2	0.176
W4	1	1	1	3	0
W5	1	1	0	2	0.176
W6	0	1	0	1	0.477
W7	0	1	0	1	0.477
W8	0	0	1	1	0.477
W9	0	0	1	1	0.477
W10	0	0	1	1	0.477

- TF-IDF
  - D1 = (0.477,0.477,0.176,0,0.176,0,0,0,0,0)
  - D2 = (0,0,0,0,0.176,0.477,0.477, 0,0,0)
  - D3 = (0,0,0.176,0,0,0,0,0.477,0.477, 0.477)
- Normalization
  - D1= [0.6634,0.6634,0.2448,0,0.2448,0,0,0,0,0]
  - D2= [0,0,0,0,0.2525,0.6842,0.6842,0,0,0]
  - D3= [0,0,0.2084,0,0,0,0,0.5647,0.5647,0.5647]

Query

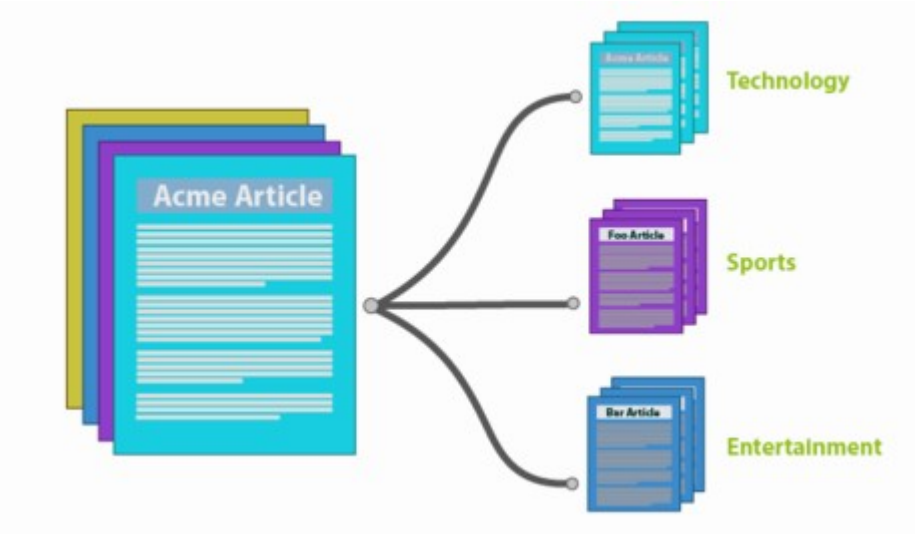
- Q=(0,0,0,0,0.176,0,0,0,0,0)
- Normalization: (0,0,0,0,1,0,0,0,0,0)

- Cosine_sim(Q,D1)=0.2448
- Cosine_sim(Q,D2)=0.2525
- Cosine_sim(Q,D3)=0

The most relevant

Other text mining tasks

- Documents clustering
- Documents classification
- In these tasks, we represent each document using the same **vector-space model**
- Classification methods and clustering methods can then be used



However…

- Oxford Dictionary has 273,000 headwords; 171,476 of them being in current use, 47,156 being obsolete words and around 9,500 derivative words included as subentries.
- Note that the stopping word list currently contains less than 1000 words.
- The vector may be tooooooooo long even if we delete all the stop words
- Dimension reduction is highly deserved

Dimension reduction for text*

- Latent Semantic Indexing
  - Doing Singular Value Decomposition (SVD) on the matrix, each row represents a document, each column represents one word's corresponding TF-IDF (or normalized TF-IDF)
  - Keep the **left** singular vectors corresponding to first d largest singular values-the dimensionality is reduced to d
  - The matrix is sparse: the non-zero entries in the matrix is far less than the size of the matrix (Python implementation for sparse SVD)
- Locality Preserving Indexing
  - Keep the local similarity between documents
- …

Summary

- Text retrieval
  - Given the query, retrieve the relevant documents
  - Represent the query and each document by the vector-space model by calculating the TF-IDF
  - Calculate the cosine similarity between the query and each document
  - Rank the documents' relevance according to the cosine similarity
- Text classification/clustering
  - Represent each document by the same vector-space model
  - Dimension reduction is highly deserved

Recommended reading

- Please refer to links inline

Next week

- Mining other forms of data rather than text
- Final-term review