



Trump's Taiwan Call

What are people tweeting regarding Trump's Taiwan call ?

Dataset Pitch (Team 4)

趙上涵 Ricardo S. Chao

楊曜丞 Dylan Yang

趙晟瑋 Spencer Chao

Outline



CONCLUSION



PROBLEM
DOMAIN



DATA OVERVIEW



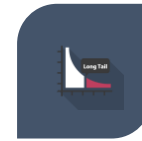
NOISE
ESTIMATING &
DATA SPLITTING



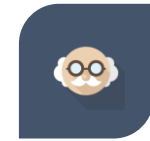
DATA
REPRESENTATION



INTERPRETATION



MODELING



CONCLUSION

Problem Domain



CONCLUSION



PROBLEM
DOMAIN



DATA OVERVIEW



NOISE
ESTIMATING &
DATA SPLITTING



DATA
REPRESENTATION



INTERPRETATION



MODELING



CONCLUSION

PROBLEM DOMAIN: Issue

“We are curious about the political tendency of each tweet. Could we classify all the tweets into different groups with different political tendencies?”



Pro-Trump



Anti-Trump

PROBLEM DOMAIN: Goal

“To resolve the issue, our main goal is to apply unsupervised algorithm we’ve learned to our dataset (tweets) with tools manufactured by ourselves and with mathematical concept in our mind.”



Mathematical Concepts



More Handmade
Models



Natural Language
Processing Techniques

Data Overview



CONCLUSION



PROBLEM
DOMAIN



DATA OVERVIEW



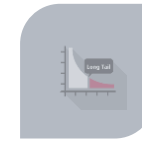
NOISE
ESTIMATING &
DATA SPLITTING



DATA
REPRESENTATION



INTERPRETATION



MODELING



CONCLUSION

Dataset Overview

kaggle.com

≡ kaggle

🔍 Search

- 🏠 Home
- 🏆 Compete
- 📊 Data
- 📓 Notebooks
- 💬 Discuss
- 🎓 Courses
- ⋮ More

Recently Viewed

- 📊 Trump's Taiwan Call
- 🎬 IMDB Movies Dataset
- 👤 The reddit self-post cl...
- 🌳 Number of Taiwan Airp...
- 🧠 TReNDS Neuroimaging

📁 Dataset

^ 2

Trump's Taiwan Call

What are people tweeting regarding trump's taiwan call ?



Ujjwal Saxena • updated 3 years ago (Version 1)

[Data](#) [Tasks](#) [Kernels](#) [Discussion](#) [Activity](#) [Metadata](#)

Download (252 KB)

[New Notebook](#)

🏠 Usability 6.3

⚖️ License Database: Open Database, Contents: Database Contents

🏷️ Tags news

Description

This data set contains recent tweets regarding The trump's call to Taiwan.
I mined twitter for recent 2000 tweets. it's 3rd Dec, 2016, 6:19 Indian Standard Time.
Can be used for sentiment analysis and future predictions.
Thanx to me and Twitter
Did You like it ??

Dataset Overview

TWEETS

RT @SteveSGoddard: The @nytimes was thrilled by Obama talking to Communist dictator Castro, and horrified by Trump talking to the elected I??e

Oops. RT @thehill: Trump breaks decades of US protocol with phone call to Taiwan leader <https://t.co/YU2x04PhV7> <https://t.co/2WXDViC7IK>

Donald Trump? call with Taiwan president was no surprise: Official <https://t.co/F8TzF1U9UPps>

RT @ShubieDobee: So, Trump spoke to the president of Taiwan.

Trumps wants to build hotel and resorts in Taiwan.

Does anyone see a CONFLICT??n

Same mental midgets that defended Castro <https://t.co/Nrt2yCb7M0> are mad Trump talked to Taiwan

RT @guardian: Trump's Taiwan phone call preceded by hotel development inquiry <https://t.co/nqKTBVr3eK>

RT @laseptiemewilay: 1-About Trump's phone call to Taiwan's. First, it's a major diplomatic incident. Second, as part of the US normalizati??e

Trump Speaks With Taiwan? Leader, an Affront to China <https://t.co/AEeTzcpNS7> #SouthChinaSea #AsiaPacificnd

RT @RawStory: REVEALED: Trump Organization eyes Taiwan for newest luxury resort location <https://t.co/RuaFG6YTub> <https://t.co/je0TE0LXbE>

RT @terrymendozer: Obama will not release info to congress on refugees deal and y'all worried about a congratulatory call from our ally Tai??e

RT @thehill: Trump says Taiwan leader called him, though Taiwan paper says Trump staffer arranged call <https://t.co/QYprHM5BLs> <https://t.co/??e>

RT @_Makada_: Hoaxing media says Trump isn't allowed to talk to Taiwan, but Obama can sell them \$1.83 billion worth of weapons?

<https://t.??>:

RT @kurteichenwald: I wonder how conservs will respond to Trump having just reversed a China/Taiwan policy put in place by Reagan & honored??

RT @realwesmyers: Can you believe Donald Trump took a call from Taiwan without having them donate to his private foundation first?

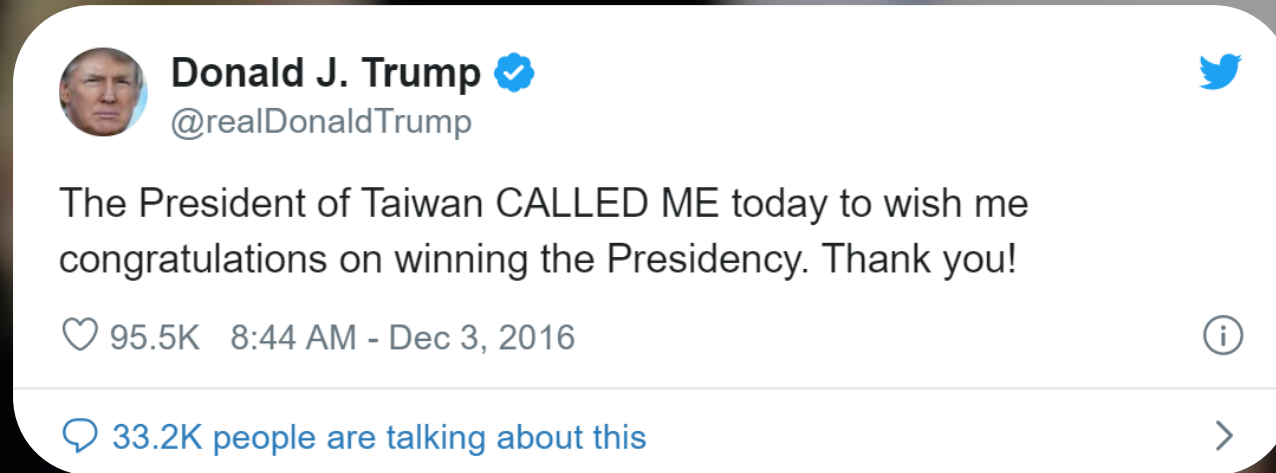
RT @terrymendozer: Obama will not release info to congress on refugees deal and y'all worried about a congratulatory call from our ally Tai??e

@MarkJKremer @realDonaldTrump MSM continues to lie lie lie. Taiwan called Trump. He didn't call Taiwan.

RT @EricLiptonNYT: Trump Org: "we are not planning any expansion into Taiwan." How abt this? Great find @mega2e <https://t.co/zBLAHLeJmz> htt??e



Dataset Overview



Trump's phone call with Taiwan president risks China's wrath

The Guardian | Sat 3 Dec 2016 10.19 GMT | <https://reurl.cc/Y14aoL>

Donald Trump looked to have sparked a potentially damaging diplomatic row with China on Friday after speaking to Taiwanese president Tsai Ing-wen on the telephone in a move experts said would anger Beijing.

The call, first reported by the Taipei Times and confirmed by the Financial Times, is thought to be the first between the leader of the island and a US president or president-elect since ties between America and Taiwan were severed in 1979, at Beijing's behest.

Dataset Overview

RT @SteveSGoddard:

The @nytimes was thrilled by Obama talking to Communist dictator Castro, and horrified by Trump talking to the elected I??e



Retweet Without Comment

@Canvasguru: Lol!

RT @martintagg: .@realDonaldTrump You've been popular with ladies for a while. What's your secret handsome?" My hair!



Retweet With Comment

Kate is donating a #kidney to her husband <http://t.co/Xg95n9tFkE> . You can help! I did @fundanything #donate




Tweet

Dataset Overview

Tweet Type **At Sign**

RT **@SteveSGoddard:** **Retweet Text**

The @nytimes was thrilled by Obama talking to Communist dictator Castro, and horrified by Trump talking to the elected I??e



Retweet Without Comment

@Canvasguru: **Lol!** **Personal Comment**


RT @martintagg: .@realDonaldTrump You've been popular with ladies for a while. What's your secret handsome? My hair!



Retweet With Comment

Tweet Content

Kate is donating a #kidney to her husband **http://t.co/Xg95n9tFkE** . You can help! I did @fundanything **#donate**



Tweet

Dataset Overview

- 1229 tweets. (with 775 unique values.)

(May have to crawl additional data from twitter for later word2vec model build-up)

- .csv file (each row represent an entry of the data.)
- Include various “Twitter Slangs”, such as @ Sign, Hashtag , RT, etc.
- Lots of noise, and misspellings.
- Without labels.



Dataset Overview

Mojibake

Donald Trump? 雜? call with Taiwan
president was no surprise: Official
<https://t.co/F8TzF1U9UPps>



RT @terrymendozer: Obama will not
release info to congress on refugees
deal and y'all worried about a
congratulatory call from our ally

Tai??e

Misspelling



RT @charliekirk11: Trump takes a
congratulatory phone call from Taiwan
and gets from grief from the media
than Obama giving \$500 billion to??e



Misspelling

→ Before applying models, **noise** should be estimated.

Noise estimating & Data splitting



CONCLUSION



PROBLEM
DOMAIN



DATA OVERVIEW



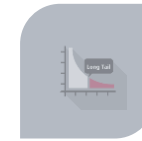
NOISE
ESTIMATING &
DATA SPLITTING



DATA
REPRESENTATION



INTERPRETATION



MODELING



CONCLUSION

Purpose



Noise estimating

To make sure that only the important part of the corpora are fed into the models and ensure the accuracy of the result.

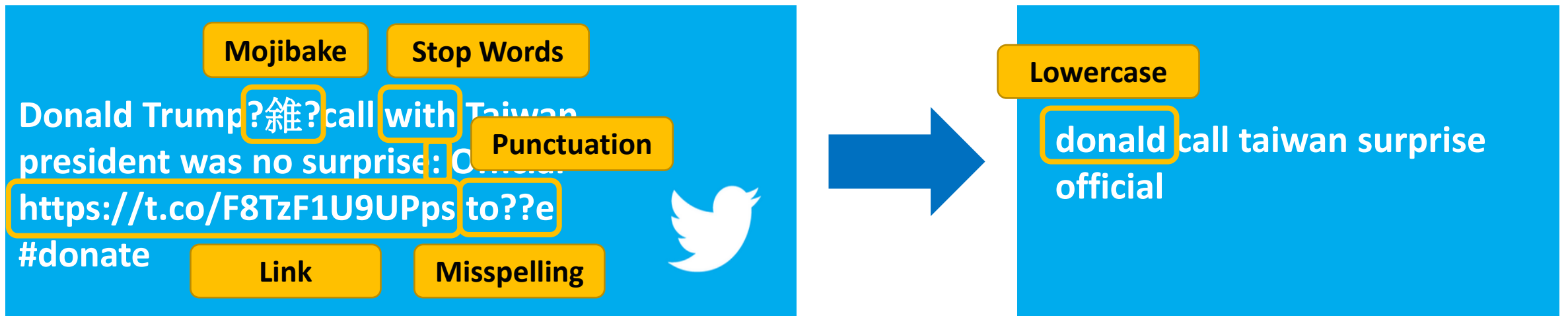


Data splitting

To split the important content of the tweet into individual vocabularies and input into the model.

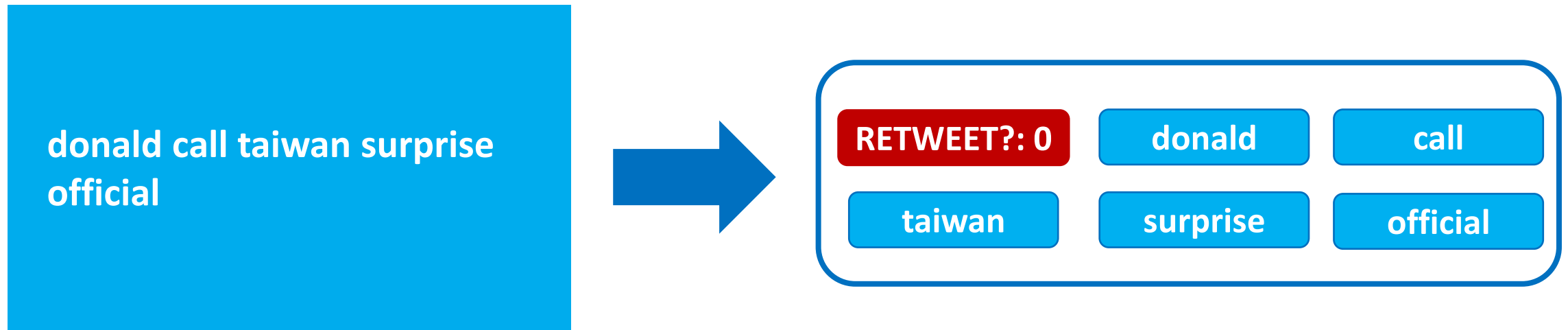
Noise estimating

To make sure that only the important part of the corpora are fed into the models and ensure the accuracy of the result. We had defined the noises:



Data splitting

To split the important content of the tweet into individual elements and input to the model.



Data splitting

is_RT	retweet_from	stopwords_removed_original_text
0		mysticwolf12001 cmon one beating clinton
0		chadwilliams91 ill move states you are hired thanks
0		hunterbalthazor ran president vote back track
0		kyleraccio vinnytitone think hell lead polls like healthy trump
0		harrycraig96
0		americanwoman00 great greta awesome thank you
0		people proud fact obama birth certificate book miraculously
0		garrykennedy great greta huge fan really inspire me thank you
0		carlotalorenabr fact realdonaldtrump run excited
1	DanScavino	join amp vpeotus pence live west allis wisconsin thankyou2016
1	DanScavino	join live mobile alabama via thankyou2016 watch
1		importing homeland agree mr trump
1	realDonaldTrump	happy
1	Reince	happy year blessings all looking forward things 2017 realdonaldtrump
1	IvankaTrump	2016 one exciting years life wish peace love happy new
1	jamestaranto	serious people worry start war trump phone taiwans crazy oe
1		photos show public
1	DanScavino	one think photo says all
1	King_Pepp	driving tons ugly realdonaldtrump talking
1	realDonaldTrump	secret getting getting started twain



Data representation



CONCLUSION



PROBLEM
DOMAIN



DATA OVERVIEW



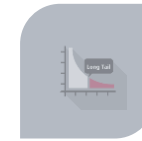
NOISE
ESTIMATING &
DATA SPLITTING



DATA
REPRESENTATION



INTERPRETATION



MODELING



CONCLUSION

Vector Space model

Since the content of tweets is mostly unstructured, the Vector Space model must be used to convert it into structured data and obtain useful information.



BOW
(BAG OF WORD)

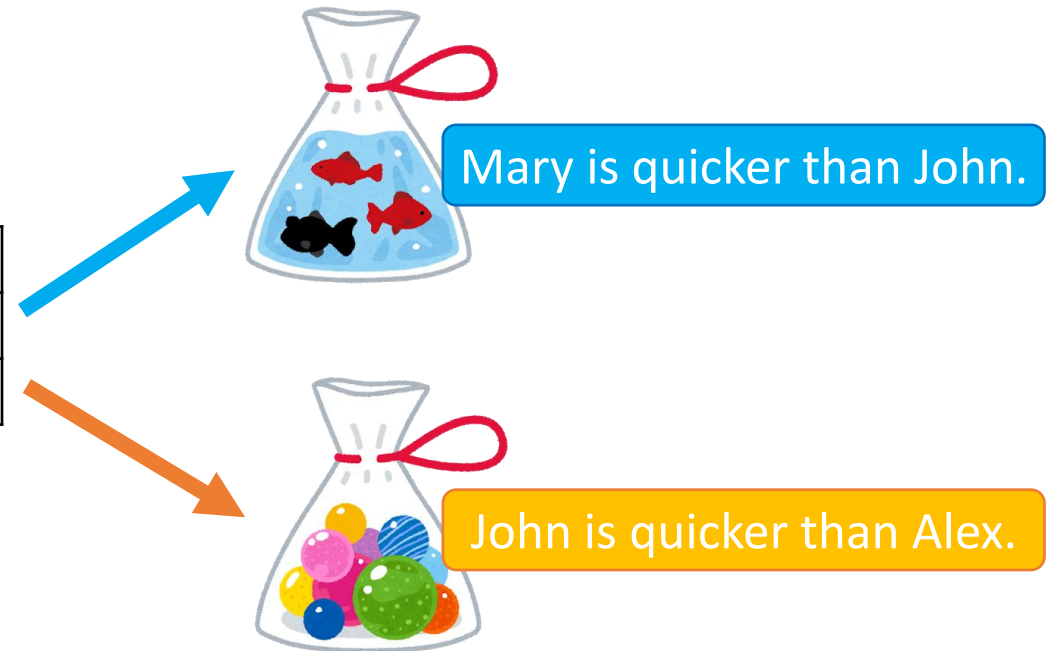


WORD2VEC

BOW (Bag of Words)

BOW is a common representation of documents. The representation of a document d is the set of weight of its term.

Sentence	Mary	John	Alex	quicker
Mary is quicker than John.	1	1	0	1
John is quicker than Alex.	0	1	1	1



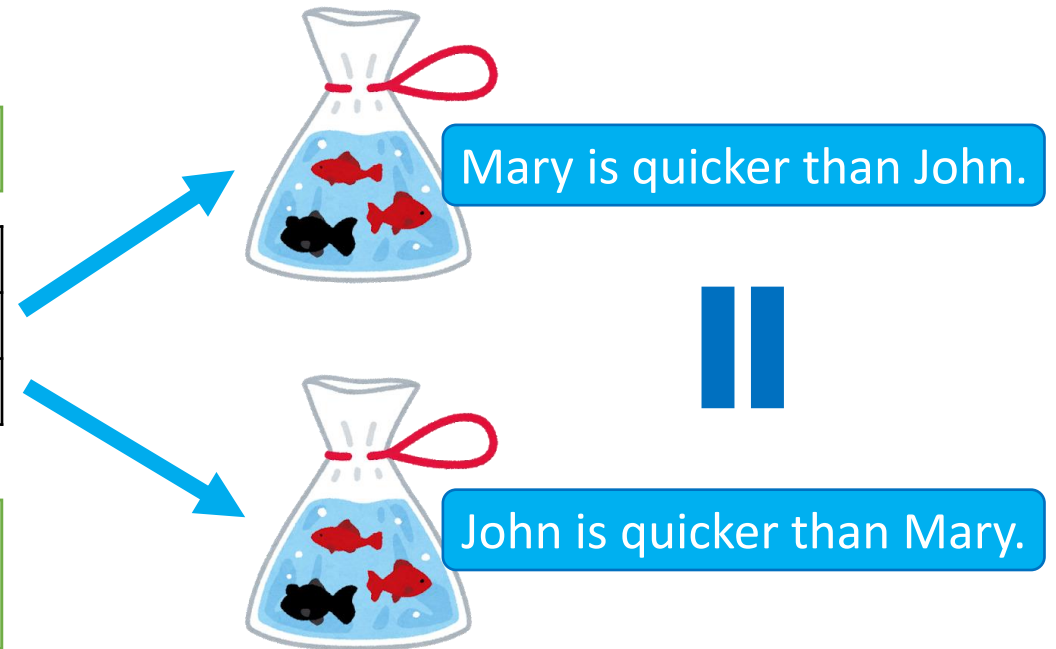
BOW (Bag of Words)

BOW is a common representation of documents. The representation of a document d is the set of weight of its term.

[CONS. 1] The **Ordering** of the terms is ignored!!!

Sentence	Mary	John	Alex	quicker
Mary is quicker than John.	1	1	0	1
John is quicker than Mary.	1	1	0	1

[CONS. 2] If most of the feature vectors are sparse, bag-of-words feature matrix would be a **sparse matrix**!!!



TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is an unsupervised learning algorithm and a statistical method that often be used to evaluate how important a word is to whole data. **TF-IDF Keyword extraction** is commonly used as a **weighted** method of information retrieval and text exploration.

TF

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

The frequency of the word's occurrence



IDF

$$idf_i = \log \frac{|D|}{|\{j: t_i \in dj\}|}$$

Inverse Document Frequency



TF-IDF

The importance of the word

TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is an unsupervised learning algorithm and a statistical method that often be used to evaluate how important a word is to whole data. **TF-IDF Keyword extraction** is commonly used as a **weighted** method of information retrieval and text exploration.

The weight of term t in document d is:

High, when t occurs many times in d and appears within a small number of documents.

Low, when t is a rare term in d and occurs in virtually all documents in the collection.

Sentence	Mary	John	Alex	quicker
Mary is quicker than John.	0.34	1.56	0	5.16
John is quicker than Alex.	0	1.56	0.54	5.16

