DataLab Cup 1: Predicting News Popularity

DataLab

Outline

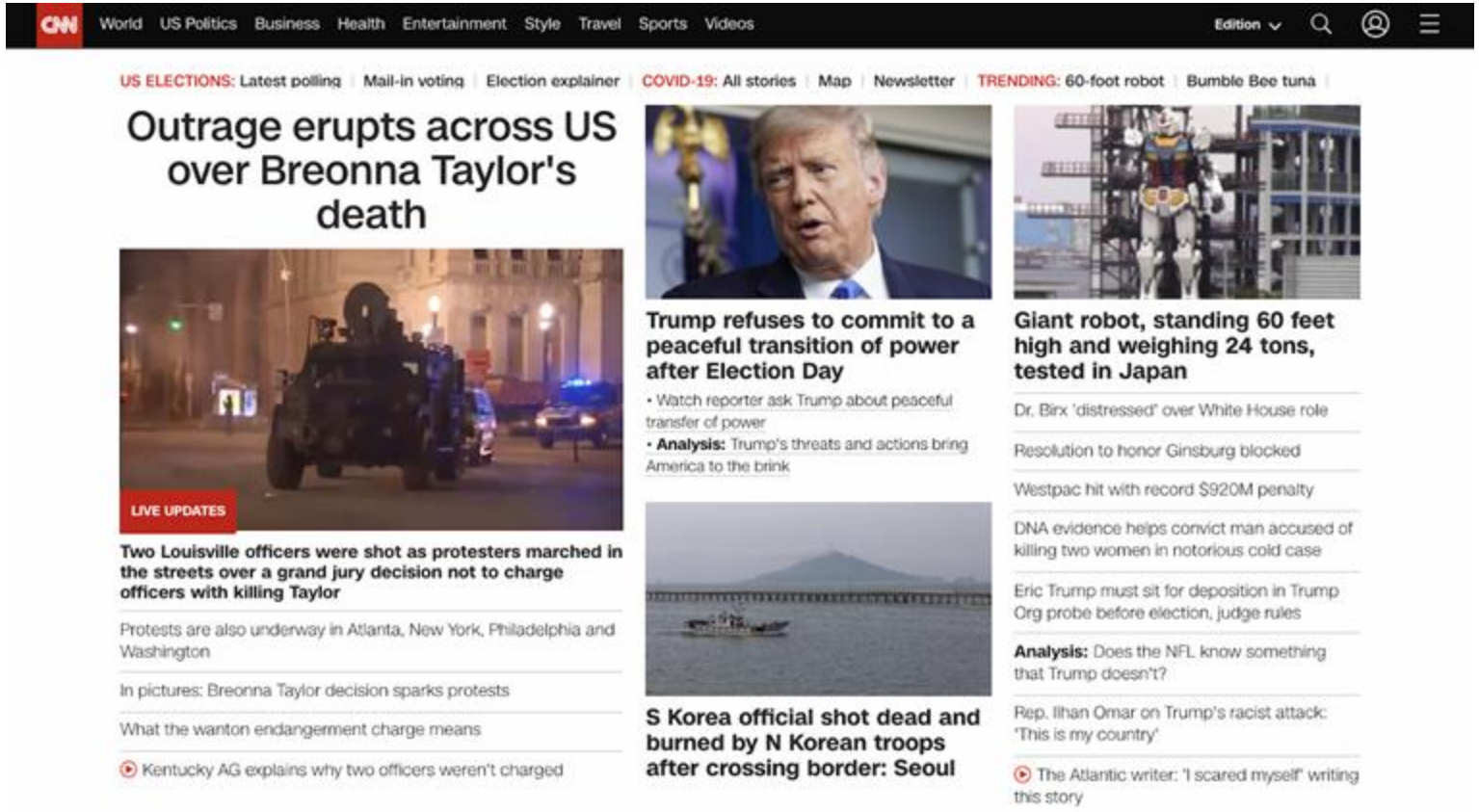
- Competition information
- Method
- Evaluation metric
- Report & Show off
- Precautions

Outline

- Competition information
- Method
- Evaluation metric
- Report & Show off
- Precautions

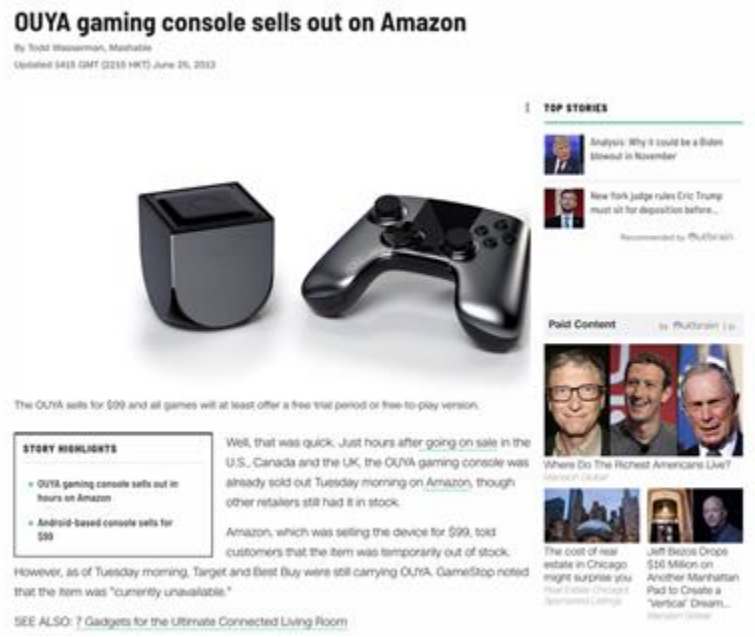
Competition Information - Goal

- Predict the news popularity



Competition Information - Dataset

- Training data
  - 27643 pieces of news
- Testing data
  - 11847 pieces of news



Popularity (1/-1)


```
<html><head><div class="article-info"><span class="byline "><a href="/author/todd-wasserman/"><img alt="2016%2f06%2f30%2f5f%2f20130712e9toddwasserm.01447.37a4f" class="author_image" src="http://i.amz.mshcdn.com/fcra-FTS5Tjwno2uuh7HD56bIgw=/90x90/2016%2F06%2F30%2F5f%2F20130712e9toddwasserm.01447.37a4f.jpg"/></a><span class="author_name">By <a href="/author/todd-wasserman/">Todd Wasserman</a></span><time datetime="Tue, 25 Jun 2013 12:54:54 +0000">2013-06-25 12:54:54 UTC</time></span></div></head><body><h1 class="title">OUYA Gaming Console Already Sold Out on Amazon</h1><figure class="article-image"><img class="microcontent" data-fragment="lead-image" data-image="http://i.amz.mshcdn.com/_nNuKEPHjD0JuHdMnX0iZEt7nok=/950x534/2013%2F06%2F25%2F1a%2FOUYA.3d2cc.jpg" data-micro="1" data-url="null" src="http://i.amz.mshcdn.com/_nNuKEPHjD0JuHdMnX0iZEt7nok=/950x534/2013%2F06%2F25%2F1a%2FOUYA.3d2cc.jpg"/></figure><article data-channel="business"><section class="article-content"> <p>Well, that was quick. Just hours after <a href="http://mashable.com/2013/06/25/ouya-launch-day/">going on sale</a> in the U.S., Canada and the UK, the OUYA gaming console was already sold out Tuesday morning on <a href="http://mashable.com/category/amazon/">Amazon</a>, though other retailers still had it in stock. </p> <p>Amazon, which was selling the device for $99, told customers that the item was <a href="http://www.amazon.com/OUYA-Console/dp/B0050SZD18/ref=sr_tr_sr_1?ie=UTF8&amp;qid=1372163367&amp;sr=8-1&amp;keywords=ouya" target="_blank">temporarily out of stock</a>. However, as of Tuesday morning, Target and Best Buy were still carrying OUYA. GameStop noted that the item was "<a href="http://www.gamestop.com/elect/consoles/ouya-game-console/107853" target="_blank">currently unavailable</a>."</p> <div class="see-also"><p><strong>SEE ALSO: <a href="http://mashable.com/2013/06/17/connected-living-room/">7 Gadgets for the Ultimate Connected Living Room</a></strong></p></div> <p>OUYA launched on <a href="http://mashable.com/category/kickstarter/">Kickstarter</a> as an open gaming console that anyone could develop for or hack as they see fit, all for a $99 price tag. The Kickstarter hit its $900,000 funding goal in eight hours, and broke Kickstarter records after raising $8.6 million total. Earlier this year, OUYA's creators announced that the console would be widely available at retail stores in June. </p> <p>Kickstarter backers, meanwhile, began receiving their OUYA consoles in April.</p> <p><em>Image courtesy of <a href="http://tektab.com/" target="_blank">Saad Faruque</a></em></p> </section></article><footer class="article-topics"> Topics: <a href="/category/amazon/">amazon</a>, <a href="/category/amazon-kindle/">amazon kindle</a>, <a href="/category/business/">Business</a>, <a href="/category/gaming/">Gaming</a> </footer></body></html>
```

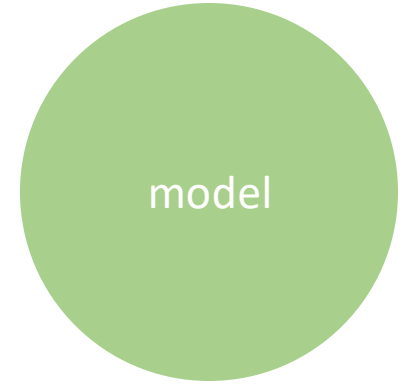
Outline

- Competition information
- Method
- Evaluation metric
- Report & Show off
- Precautions

Method - Feature Engineering

- The act of extracting features from raw data and then transforming them into something that we can use for a machine learning model.

<html><head> By Todd Wasserman<time datetime="Tue, 25 Jun 2013 12:54:54 +0000">2013-06-25 12:54:54 UTC</time> </head><body><h1 class="title">OUYA Gaming Console Already Sold Out on Amazon</h1><figure class="article-image"></figure><article data-channel="business"><section class="article-content"> Well, that was quick. Just hours after going on sale in the U.S., Canada and the UK, the OUYA gaming console was already sold out Tuesday morning on Amazon, though other retailers still had it in stock. Amazon, which was selling the device for $99, told customers that the item was temporarily out of stock. However, as of Tuesday morning, Target and Best Buy were still carrying OUYA. GameStop noted that the item was "currently unavailable." SEE ALSO: 7 Gadgets for the Ultimate Connected Living Room OUYA launched on Kickstarter as an open gaming console that anyone could develop for or hack as they see fit, all for a $99 price tag. The Kickstarter hit its $900,000 funding goal in eight hours, and broke Kickstarter records after raising $8.6 million total. Earlier this year, OUYA's creators announced that the console would be widely available at retail stores in June. Kickstarter backers, meanwhile, began receiving their OUYA consoles in April. Image courtesy of Saad Faruque </section></article><footer class="article-topics"> Topics: amazon, amazon kindle, Business, Gaming </footer></body></html>



Raw Data

Feature Engineering

Classification

Method - Feature Engineering

- Data preprocessing
- Convert words to vectors

Method - Feature Engineering

- Data preprocessing
- Convert words to vectors

Data Preprocessing

- Process HTML tags
  - Install Beautiful Soup to remove HTML tags
  - Select the specific tags

By Todd Wasserman2013-06-25 12:54:54 UTCOUYA Gaming Console Already Sold Out on Amazon Well, that was quick. Just hours after going on sale in the U.S., Canada and the UK, the OUYA gaming console was already sold out Tuesday morning on Amazon, though other retailers still had it in stock. Amazon, which was selling the device for $99, told customers that the item was temporarily out of stock. However, as of Tuesday morning, Target and Best Buy were still carrying OUYA. GameStop noted that the item was "currently unavailable." SEE ALSO: 7 Gadgets for the Ultimate Connected Living Room OUYA launched on Kickstarter as an open gaming console that anyone could develop for or hack as they see fit, all for a $99 price tag. The Kickstarter hit its $900,000 funding goal in eight hours, and broke Kickstarter records after raising $8.6 million total. Earlier this year, OUYA's creators announced that the console would be widely available at retail stores in June. Kickstarter backers, meanwhile, began receiving their OUYA consoles in April. Image courtesy of Saad Faruque Topics: amazon, amazon kindle, Business, Gaming

Data Preprocessing

- Process HTML tags
  - Install Beautiful Soup to remove HTML tags
  - Select the specific tags

<html><head> By Todd Wasserman<time datetime="Tue, 25 Jun 2013 12:54:54 +0000">2013-06-25 12:54:54 UTC</time> </head><body><h1 class="title">OUYA Gaming Console Already Sold Out on Amazon</h1><figure class="article-image"></figure><article data-channel="business"><section class="article-content"> Well, that was quick. Just hours after going on sale in the U.S., Canada and the UK, the OUYA gaming console was already sold out Tuesday morning on Amazon, though other retailers still had it in stock. Amazon, which was selling the device for $99, told customers that the item was temporarily out of stock. However, as of Tuesday morning, Target and Best Buy were still carrying OUYA. GameStop noted that the item was "currently unavailable." SEE ALSO: 7 Gadgets for the Ultimate Connected Living Room OUYA launched on Kickstarter as an open gaming console that anyone could develop for or hack as they see fit, all for a $99 price tag. The Kickstarter hit its $900,000 funding goal in eight hours, and broke Kickstarter records after raising $8.6 million total. Earlier this year, OUYA's creators announced that the console would be widely available at retail stores in June. Kickstarter backers, meanwhile, began receiving their OUYA consoles in April. Image courtesy of Saad Faruque </section></article><footer class="article-topics"> Topics: amazon, amazon kindle, Business, Gaming </footer></body></html>

Extract <h1>

```
<h1 class="title">OUYA Gaming Console Already Sold Out on Amazon</h1>
```

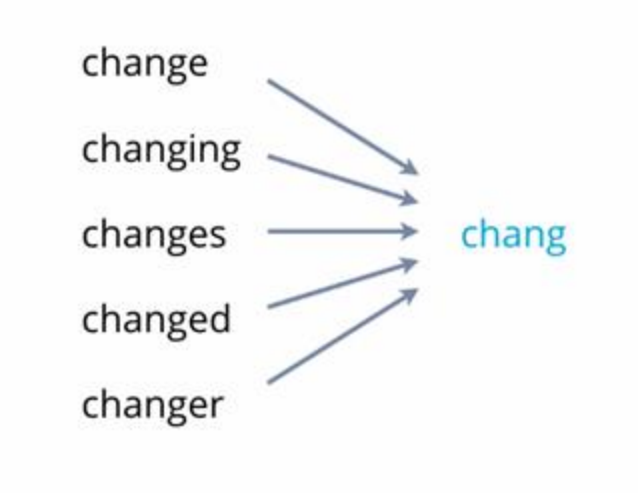
Data Preprocessing

- Stop words
  - Stop words are common in a document but less information, they might misleading the classification model.
  - You can use the stop words NLTK provided or create your own stop words.

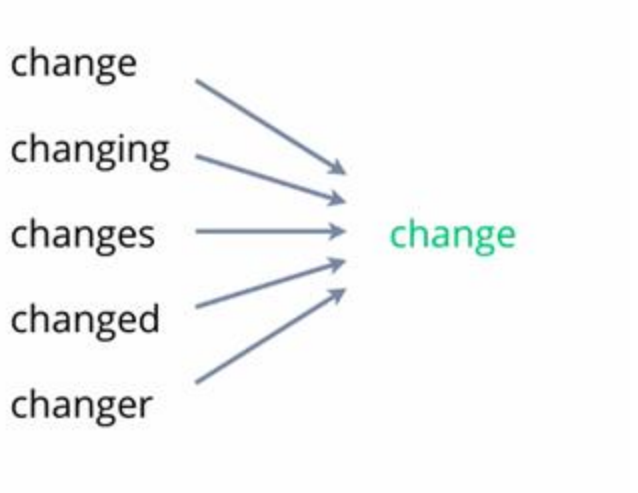
```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your',
'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "i
t's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'tha
t', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'havi
ng', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'o
f', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'abov
e', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'onc
e', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'so
me', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just',
'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'could
n', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn',
"isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "should
n't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

Data Preprocessing

- Word stemming
  - The process of reducing inflected (or sometimes derived) words to their word stem, e.g. runs, running, ran => run.
  - NLTK provides two kind of algorithms to reduce the inflected words.



Porter Stemmer



Lemmatization

Feature Engineering

- Data preprocessing
- Convert words to vectors

Convert Words to Vectors

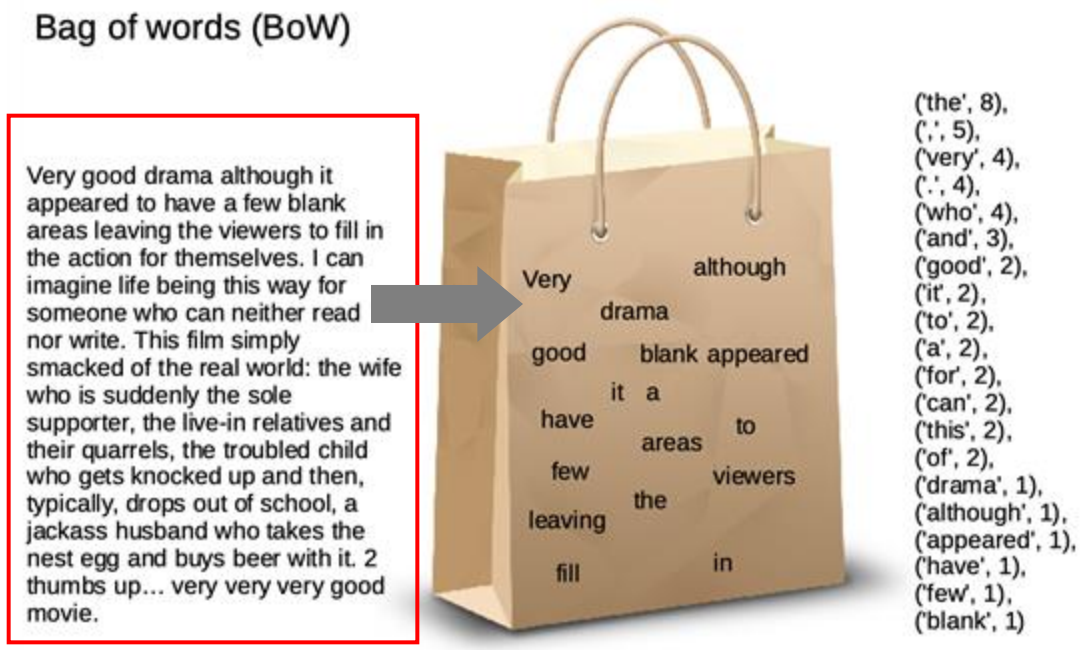
- Bag of words (BOW)
- Term frequency-inverse document frequency (TF-IDF)
- Feature hashing

Convert Words to Vectors

- Bag of words (BOW)
- Term frequency-inverse document frequency (TF-IDF)
- Feature hashing

Bag of Words

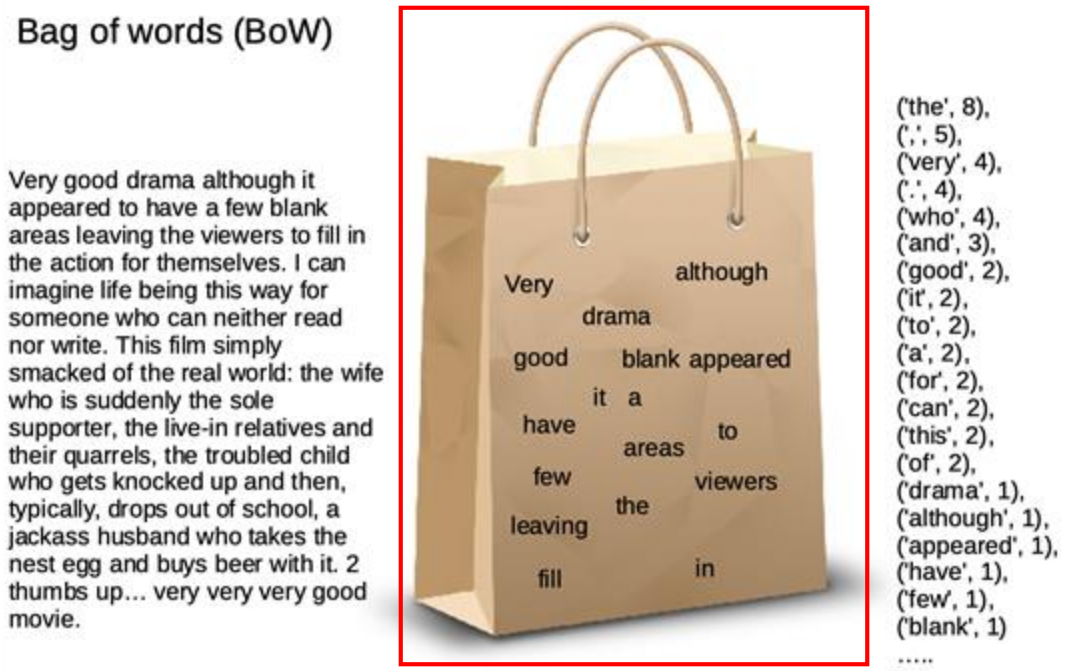
- A simple way of representing text data used in natural language processing and information retrieval.



Document

Bag of Words

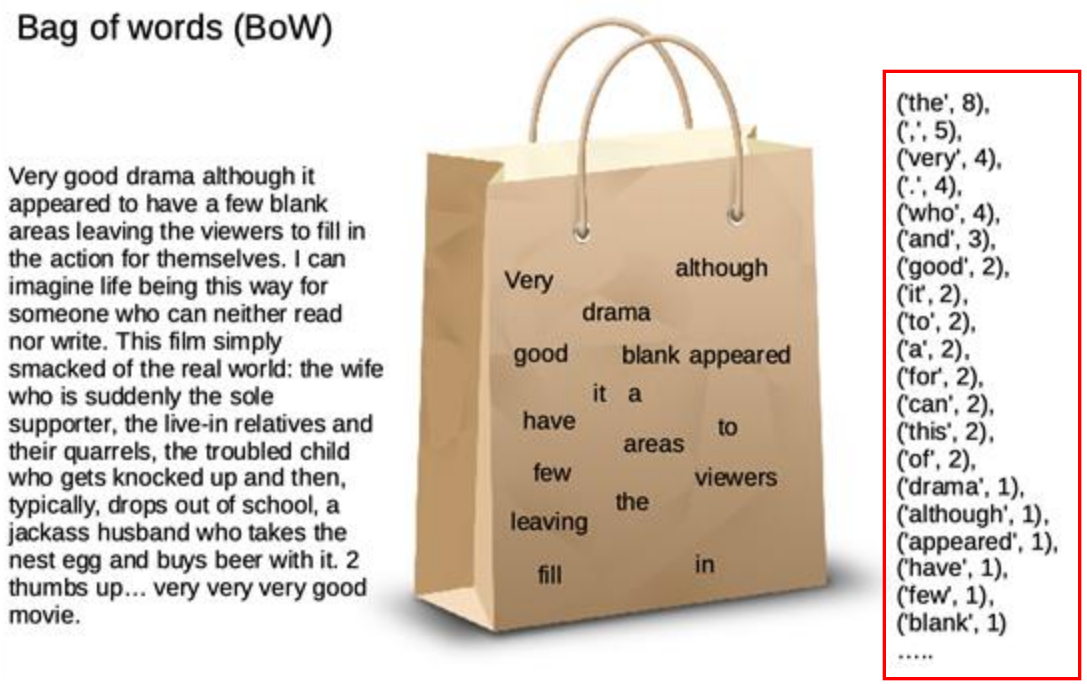
- A simple way of representing text data used in natural language processing and information retrieval



Throw all the words into the bag

Bag of Words

- A simple way of representing text data used in natural language processing and information retrieval



Record how many times each word appear in the document

Bag of Words

- Example

Review 1: This movie is very scary and long
Review 2: This movie is not scary and is slow
Review 3: This movie is spooky and good

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

Vector of Review 1: [1 1 1 1 1 1 1 0 0 0 0]

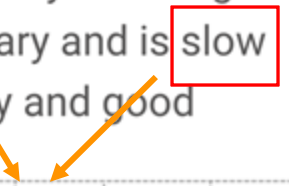
Vector of Review 2: [1 1 2 0 0 1 1 0 1 0 0]

Vector of Review 3: [1 1 1 0 0 0 1 0 0 1 1]

Bag of Words

- Example

Review 1: This movie is very scary and long
Review 2: This movie is not scary and is slow
Review 3: This movie is spooky and good



	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

Vector of Review 1: [1 1 1 1 1 1 1 0 0 0 0]

Vector of Review 2: [1 1 2 0 0 1 1 0 1 0 0]

Vector of Review 3: [1 1 1 0 0 0 1 0 0 1 1]

Bag of Words

- Example

Review 1: This movie is very scary and long
Review 2: This movie is not scary and is slow
Review 3: This movie is spooky and good

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

Vector of Review 1: [1 1 1 1 1 1 1 0 0 0 0]

Vector of Review 2: [1 1 2 0 0 1 1 0 1 0 0]

Vector of Review 3: [1 1 1 0 0 0 1 0 0 1 1]

Convert Words to Vectors

- Bag of words (BOW)
- Term frequency-inverse document frequency (TF-IDF)
- Feature hashing

TF-IDF

- A numerical statistic that intended to reflect how important a word is to a document in a collection or corpus.

- The tf-idf value increases proportionally to the number of times a word appears in a document and is offset by the number of documents in the corpus that contain the word.

TF-IDF

- Term frequency (tf)
  - It is a measure of how frequently a term appears in a document
  - $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} = \frac{Number\ of\ times\ term\ i\ appears\ in\ the\ document}{Total\ number\ of\ terms\ in\ the\ document}$

- Inverse document frequency (idf)
  - It is a measure of how important a term is
  - $idf_i = \log(\frac{|D|}{|\{j: t_i \in d_i\}|}) = \log(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ i\ in\ it})$

- $tf\text{-}idf_{i,j} = tf_{i,j} * idf_i$

Convert Words to Vectors

- Bag of words (BOW)
- Term frequency-inverse document frequency (TF-IDF)
- Feature hashing

Feature Hashing

- Reduce the dimension vocabulary space by hashing each vocabulary into a hash table with a fixed number of buckets.
- There are some cons, e.g. the information will be less than TF-IDF.
- However, you can do out-of-core learning when using hashing which means you only load part of the dataset at a time similar to the concept of batch.

Outline

- Competition information
- Method
- Evaluation metric
- Report & Show off
- Precautions

Evaluation Metric

- We will use **Area Under Curve (AUC)** as our metric in this competition.
- AUC is the area under the ROC curve.

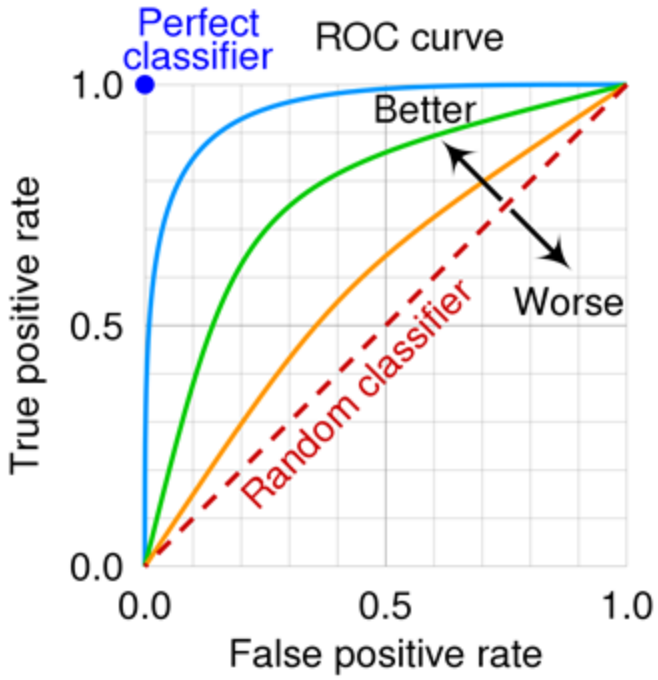
Evaluation Metric

- Confusion matrix
  - A specific table layout that allows visualization of the performance of an algorithm.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Evaluation Metric

- Receiver operating characteristic curve (ROC curve).
  - A graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.
  - False positive rate $= \frac{False\ positive}{ALl\ negative\ samples}$.
  - True positive rate $= \frac{Ture\ positive}{ALl\ positive\ samples}$.



Outline

- Competition information
- Method
- Evaluation metric
- Report & Show off
- Precautions

Report

- Each team has to hand in a report after the competition ends.
- The report should be in the form of a Jupyter Notebook.
- The filename should be DL_comp1_{student ID}_report.ipynb
- Please submit your report via the eeclass system before 2024/10/27.

Report

- The report should include:
  - Student ID & name of each member
  - How did you preprocess the data
    - E.g. cleaning, feature engineering
  - How did you build the classifier
    - E.g. model, training algorithm, special techniques
  - Conclusion
    - E.g. interesting findings, pitfalls, takeaway lessons

Show off

- For the top three scores on the private leaderboard, a show off event will be held during the class on 2024/10/24.
- For the top three students, please prepare a brief presentation (about 3 to 5 minutes) to introduce your feature design, model, methods tried and experiment results.
- Please prepare slides. Don’t present directly with your code.

Outline

- Competition information
- Method
- Evaluation metric
- Report & Show off
- Precautions

Precautions

- Timeline:
  - 2024/09/26 (Thur) - competition announced
  - 2024/10/22 (Thur) 23:59pm - competition deadline
  - 2024/10/27 (Sun) 23:59pm - report deadline (to eeclass)
  - 2024/10/24 (Thur) - show off (TOP 3)
- Scoring:
  - Private leaderboard - 80%
  - Report - 20%

Precautions

- Rules
  - What you **CAN** do
    - Use untaught APIs: you can use any machine learning tools you like as well as models/techniques that are not taught in the class.
  - What you **CAN'T** do
    - Create specific deterministic rules that make predictions.
    - Train models using representation learning based on neural networks.
    - Use datasets and references beyond those made available by the competition.
    - Abuse the competition infrastructure to gain an edge.
    - Copy code from other teams.