

# DataLab Cup 1: Predicting News Popularity

Hao-Lun Sun & DataLab

# Outline

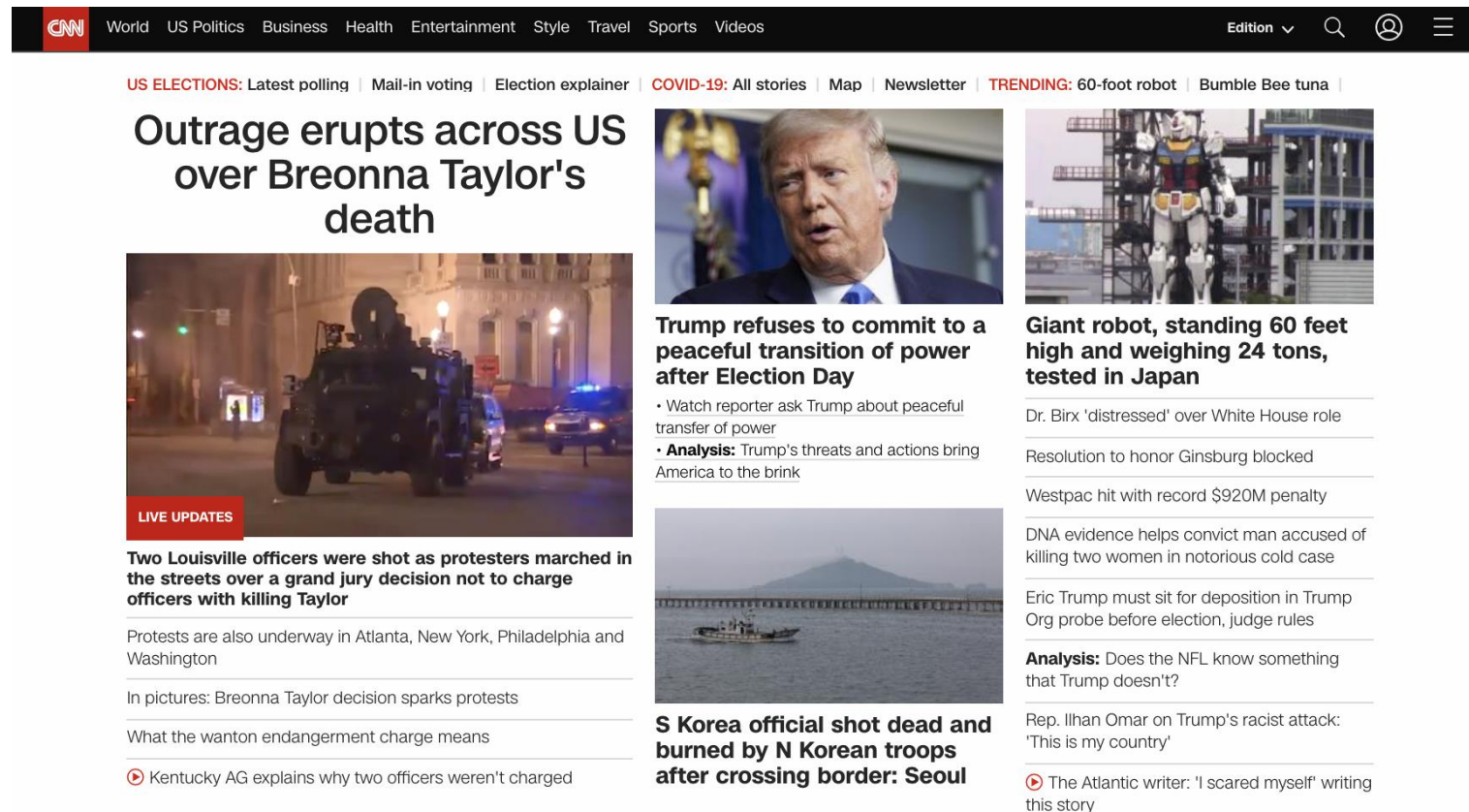
- Competition information
- Method
- Evaluation metric
- Precautions

# Outline

- Competition information
- Method
- Evaluation metric
- Precautions

# Competition Information - Goal

- Predict the news popularity



The screenshot shows the CNN website homepage. The navigation bar at the top includes the CNN logo and links to World, US Politics, Business, Health, Entertainment, Style, Travel, Sports, and Videos. There are also links for Edition, search, and user profile. Below the navigation bar, there are several news headlines and images. The main headline is 'Outrage erupts across US over Breonna Taylor's death' with a 'LIVE UPDATES' tag. Other headlines include 'Trump refuses to commit to a peaceful transition of power after Election Day', 'Giant robot, standing 60 feet high and weighing 24 tons, tested in Japan', 'S Korea official shot dead and burned by N Korean troops after crossing border: Seoul', and 'Kentucky AG explains why two officers weren't charged'. There are also smaller headlines about Dr. Birx, Resolution to honor Ginsburg, Westpac, DNA evidence, Eric Trump, and Rep. Ilhan Omar.

**Outrage erupts across US over Breonna Taylor's death**

**Trump refuses to commit to a peaceful transition of power after Election Day**

**Giant robot, standing 60 feet high and weighing 24 tons, tested in Japan**

**S Korea official shot dead and burned by N Korean troops after crossing border: Seoul**

**Kentucky AG explains why two officers weren't charged**

# Competition Information - Dataset

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

## Ouya gaming console sells out on Amazon

By Todd Wasserman, Mashable  
Updated 1415 GMT (2215 HKT) June 25, 2013



The Ouya sells for \$99 and all games will at least offer a free trial period or free-to-play version.

### STORY HIGHLIGHTS

- OUYA gaming console sells out in hours on Amazon
- Android-based console sells for \$99

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



## Popularity (1/-1)

```
<html><head><div class="article-info"><span class="byline "><a href="/author/todd-wasserman/"></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"></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
- Precautions

# Method - Feature Engineering

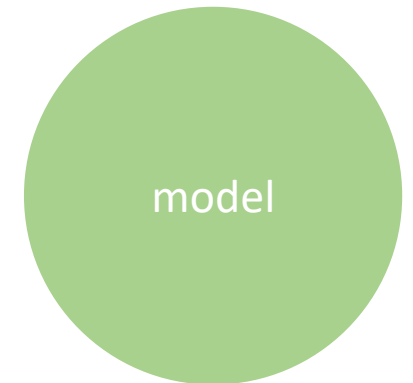
- The act of extracting features from raw data and then transforming them into something that we can use for a machine learning model.

```
<html><head><div class="article-info"><span class="byline"><a href="/author/todd-wasserman/"></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"></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/B0050S2D18/ref=sr_tr_sr_1?ie=UTF8&qid=1372163367&sr=8-l&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>
```

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

```
<html><head><div class="article-info"><span class="byline"><a href="/author/todd-wasserman/"></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></div></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"> <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/B0050S2D18/ref=sr_tr_sr_l?ie=UTF8&qid=1372163367&sr=8-1&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/">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>
```



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><div class="article-info"><span class="byline "><a href="/author/todd-wasserman/"></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"></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&qid=1372163367&sr=8-1&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>
```



Extract <h1>

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

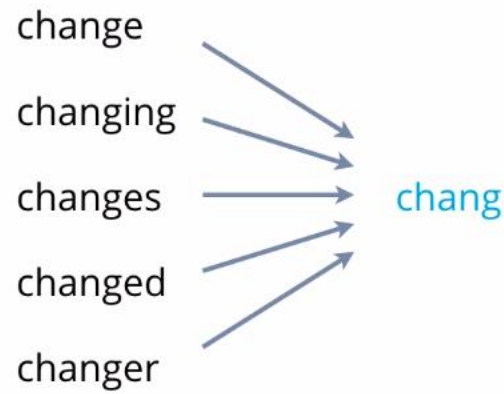
# Data Preprocessing

- 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.

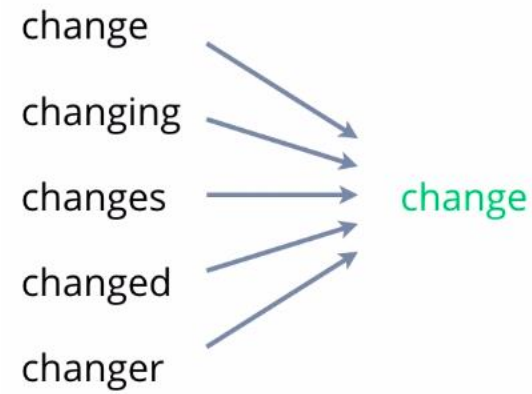
```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', '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", 'couldn', "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', "shouldn'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.

Bag of words (BoW)

Very good drama although it appeared to have a few blank areas leaving the viewers to fill in the action for themselves. I can imagine life being this way for someone who can neither read nor write. This film simply smacked of the real world: the wife who is suddenly the sole supporter, the live-in relatives and their quarrels, the troubled child who gets knocked up and then, typically, drops out of school, a jackass husband who takes the nest egg and buys beer with it. 2 thumbs up... very very very good movie.



('the', 8),  
(',', 5),  
( 'very', 4),  
( '.', 4),  
( 'who', 4),  
( 'and', 3),  
( 'good', 2),  
( 'it', 2),  
( 'to', 2),  
( 'a', 2),  
( 'for', 2),  
( 'can', 2),  
( 'this', 2),  
( 'of', 2),  
( 'drama', 1),  
( 'although', 1),  
( 'appeared', 1),  
( 'have', 1),  
( 'few', 1),  
( 'blank', 1)  
.....

Document

# Bag of Words

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

## Bag of words (BoW)

Very good drama although it appeared to have a few blank areas leaving the viewers to fill in the action for themselves. I can imagine life being this way for someone who can neither read nor write. This film simply smacked of the real world: the wife who is suddenly the sole supporter, the live-in relatives and their quarrels, the troubled child who gets knocked up and then, typically, drops out of school, a jackass husband who takes the nest egg and buys beer with it. 2 thumbs up... very very very good movie.



('the', 8),  
(',', 5),  
( 'very', 4),  
( '.', 4),  
( 'who', 4),  
( 'and', 3),  
( 'good', 2),  
( 'it', 2),  
( 'to', 2),  
( 'a', 2),  
( 'for', 2),  
( 'can', 2),  
( 'this', 2),  
( 'of', 2),  
( 'drama', 1),  
( 'although', 1),  
( 'appeared', 1),  
( 'have', 1),  
( 'few', 1),  
( 'blank', 1)  
.....

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

## Bag of words (BoW)

Very good drama although it appeared to have a few blank areas leaving the viewers to fill in the action for themselves. I can imagine life being this way for someone who can neither read nor write. This film simply smacked of the real world: the wife who is suddenly the sole supporter, the live-in relatives and their quarrels, the troubled child who gets knocked up and then, typically, drops out of school, a jackass husband who takes the nest egg and buys beer with it. 2 thumbs up... very very very good movie.



('the', 8),  
(',', 5),  
( 'very', 4),  
( '.', 4),  
( 'who', 4),  
( 'and', 3),  
( 'good', 2),  
( 'it', 2),  
( 'to', 2),  
( 'a', 2),  
( 'for', 2),  
( 'can', 2),  
( 'this', 2),  
( 'of', 2),  
( 'drama', 1),  
( 'although', 1),  
( 'appeared', 1),  
( 'have', 1),  
( 'few', 1),  
( 'blank', 1)  
.....

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

- $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} = \frac{\text{Number of times term } i \text{ appears in the document}}{\text{Total number of terms in the document}}$

- Inverse document frequency (idf)

- It is a measure of how important a term is

- $idf_i = \log\left(\frac{|D|}{|\{j: t_i \in d_j\}|}\right) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents with term } i \text{ in it}}\right)$

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

# 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**
- 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

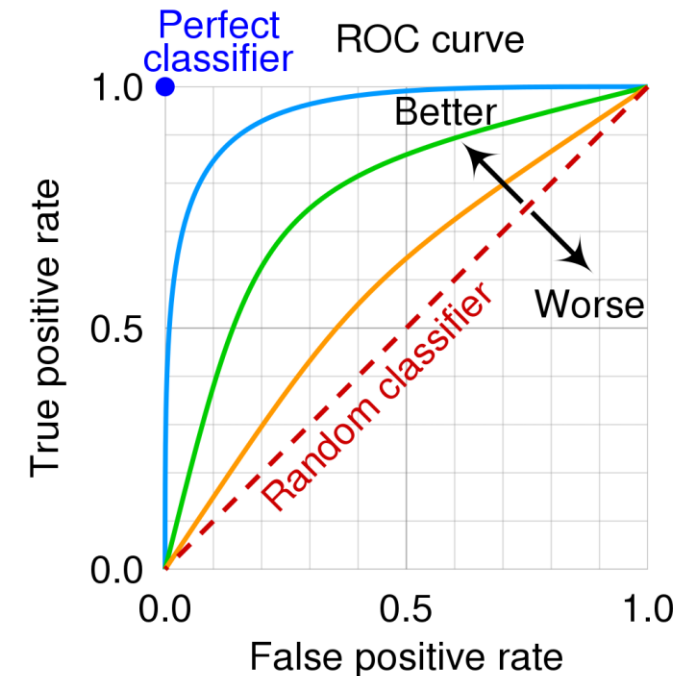
- Confusion matrix
  - A specific table layout that allows visualization of the performance of an algorithm.

	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 =  $\frac{\text{False positive}}{\text{All negative samples}}$
- True positive rate =  $\frac{\text{Ture positive}}{\text{All positive samples}}$



# Outline

- Competition information
- Method
- Evaluation metric
- Precautions



# Precautions

- Timeline:
  - 2021/10/14 (Thur) - competition announced
  - 2021/10/28 (Thur) 23:59pm (UTC) - competition deadline
  - 2021/10/31 (Sun) 23:59pm - report deadline (to eeClass)
  - 2021/11/04 (Thur) - show off
- 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.