IS590DT: Data Mining Applications

WEEK 13: Text Processing, Feature Ranking and Sentiment Analysis

Spring 2019

Yingjun Guan
yingjun2@illinois.edu
http://ischool.illinois.edu/people/yingjun-guan

ILLINOIS | School of Information Sciences
The iSchool at Illinois

Contents today

- **Introduction of text data**
- **Text processing**
- **Feature ranking**
- **TF-IDF**
- **Sentiment Analysis**
- **Summary**

- What's the first word that come to your mind when talking about "Text Mining"?
- Any differences compared with "data mining"?

Introduction of text data

- Today, we are talking about words vs docs.
- What's the relationship …
  - Among docs;
    1. Give me some definitions of distance.
    2. Why/what do we care about the docs?
    3. Compared with words, what's doc's characteristics.
    4. How hard is it to get the right doc?

Introduction of text data

- Today, we are talking about words vs docs.
- What's the relationship …
  - Among words;
    1. Give me some definitions of distance.
    2. Why/what do we care about the words?
    3. Compared with documents, what's word's characteristics.
    4. How hard is it to get the right word?

Introduction of text data

- Today, we are talking about words vs docs.
- What's the relationship …
  - Between docs and words.
    1. Can I represent the doc with words?
    2. How to simplify?
    3. Can I represent the word with doc?
    4. How to simplify?

Why is this important?

- In the world of machine reading.
- In the world of information retrieval.
- In the world of linguistic network analysis.

Our goals……

	word1	word2	word3	…
doc1				
doc2				
doc3				
doc4				
…				

- Point of embedding;
- Point of IOT;
- Point of dimensionality.

- Your comment?
  - This helps?
  - Or not?

Text processing

- Today, let's talk about how to better use word to represent the document. <IR>
- Can it?
- How to improve?
- The goal for better IR:
  - Less words (lower dimensionality)
  - Better performance (precision & recall)

How to reduce the word dimensionality?

- **Text processing**
  - **Case;**
    - **when upper case? When lower?**
    - **Other tricks. [sentence beginning; proper noun; etc.]**
    - **Is it a good idea to turn all into upper/lower?**
    - **Any other way to combine upper and lower?**
    - **Any other thoughts?**

How to reduce the word dimensionality?

- **Text processing**
  - **Punctuation;**
    - **In what cases do you need to consider the punctuation?**
    - **In most cases, we ignore punctuations. What consequences might occur?**

How to reduce the word dimensionality?

- **Text processing**
  - **Stemming;**
    - **Why stem?**

      Eg: car, cars, car's, cars' → car

      eg: am, is, are → be

How to reduce the word dimensionality?

- **Text processing**
  - **Stemming;**
    - **Stemming vs lemmatization?**
    - *Stemming* usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.
    - *Lemma (lemmatization)* usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word.

How to reduce the word dimensionality?

- **Text processing**
  - **Stemming;**
    - **different stemming algorithms.**

(F)	Rule			Example		
	SSES	→	SS	caresses	→	caress
	IES	→	I	ponies	→	poni
	SS	→	SS	caress	→	caress
	S	→		cats	→	cat

How to reduce the word dimensionality?

- **Text processing (Stemming)**

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Pop-quiz

- Are the following statements true or false?
  - Stemming increases the size of the vocabulary.
  - In a Boolean retrieval system, stemming never lowers precision.
  - In a Boolean retrieval system, stemming never lowers recall.
  - <if example needed, go to page 18>

How to reduce the word dimensionality?

- **Text processing**
  - **Stop words**
    - Common stop words.
    - Should? Shouldn't?
    - Using different stopwords for different projects.

How to evaluate the retrieval?

- The performance {precision & recall}
- Corpus: D1{SVM}, D2{SVM}, D3{SVMs}, D4{UL}; D5{Bush}; D6{Trump}; D7{Xi}.
- If you are interested in research docs, rather than political docs, try comparing using w1{SVM} and w2{SVM_stemmed} for the performance {precision & recall}.

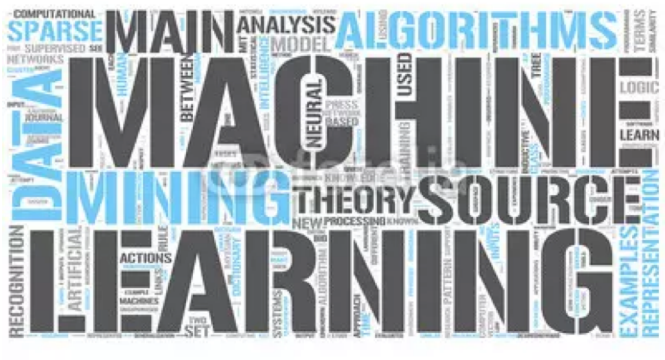
How to reduce the word dimensionality?

- **Text processing**
  - **word cloud**

Application: word cloud

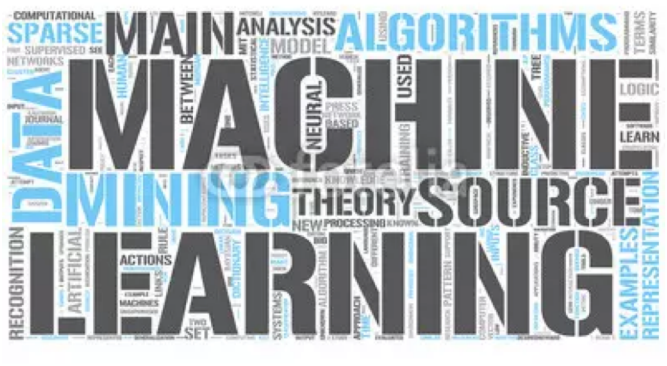
- What is word cloud?

  an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance.



- Let’s make some word cloud!
- Run your python!

What's word cloud again?



- An image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance.
- How to measure the importance?

Feature ranking

- Features (words) may weigh differently in representing the document.
- Let's see some examples first.
  - Sweet {fruit docs}
  - Apple {fruit docs}
  - Apple {technique companies}
  - Apple Inc {technique companies}

Feature ranking

- Rule of thumb standards for features.
  - Popularity
    - Enough occurrence
  - Distinctiveness
    - Able to help make prediction
  - Informativeness
    - Avoid meaningless feature {stopwords}
  - Completeness
    - Support Vector vs Vector Machine vs Support Vector Machine

Tf-idf

- tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus
- TF:term frequency, tf(*t*,*d*).

Variants of term frequency (tf) weight

weighting scheme	tf weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$

- The **inverse document frequency** is a measure of how much information the word provides, i.e., if it's common or rare across all documents

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Variants of inverse document frequency (idf) weight

weighting scheme	idf weight ($n_t = \|\{d \in D : t \in d\}\|$)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log\left(\frac{N}{1+n_t}\right)$

Example

- Corpus:
  - D1: “This is a sample, a good sample”
  - D2: “This is another sample, too.”

x	Tf(x,d1)	Tf(x,d2)	df	Tfidf(x,d1)	Tfidf(x,d2)
this					
is					
a					
good					
sample					
another					
too					

Sentiment Analysis

- Definition:
  - Computational study of opinions, sentiments, subjectivity, evaluations, attitudes, appraisal, affects, views, emotions, etc., expressed in text.
- Reference:
  - Bing Liu. Sentiment Analysis and Opinion Mining Morgan & Claypool Publishers, May 2012.

Different names.

- Many names and tasks
  - Sentiment analysis
  - Opinion mining
  - Sentiment mining
  - Subjectivity analysis
  - Affect analysis
  - Emotion detection
  - Opinion spam detection
  - Etc.

Different levels

Id: Abc123 on 5-1-2008 *"I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …"*

- One can look at this review/blog at the
  - document level, i.e., is this review + or -?
  - sentence level, i.e., is each sentence + or -?
  - entity and feature/aspect level

Polarity or not?

- Positive?
- Negative?
- In between?
- Intensity?

Word level

- Lexicon.
- Sentiment analysis tools rely on lists of words and phrases with positive and negative connotations. Many dictionaries of positive and negative opinion words were already developed. In this paper, we will look at most known words databases.

- Ref: https://medium.com/@datamonsters/sentiment-analysis-tools-overview-part-1-positive-and-negative-words-databases-ae35431a470c

Examples of lexicon

```
# -------
#
# POS    ID       PosScore          NegScore         SynsetTerms     Gloss
a        00001740          0.125    0        able#1  (usually followed by `to') having the nec
car"; "able to get a grant for the project"
a        00002098          0        0.75     unable#1        (usually followed by `to') not ha
a        00002312          0        0        dorsal#2 abaxial#1      facing away from the axis
a        00002527          0        0        ventral#2 adaxial#1     nearest to or facing towa
a        00002730          0        0        acroscopic#1    facing or on the side toward the
a        00002843          0        0        basiscopic#1    facing or on the side toward the
a        00002956          0        0        abducting#1 abducent#1  especially of muscles; dr
a        00003131          0        0        adductive#1 adducting#1 adducent#1      especiall
a        00003356          0        0        nascent#1       being born or beginning; "the nas
a        00003553          0        0        emerging#2 emergent#2   coming into existence; "a
a        00003700          0.25     0        dissilient#1    bursting open with force, as do s
a        00003829          0.25     0        parturient#2    giving birth; "a parturient heife
a        00003939          0        0        dying#1 in or associated with the process of pass
```

https://raw.githubusercontent.com/aesuli/SentiWordNet/master/data/SentiWordNet_3.0.0.txt

- Attention!
  - polysemy
  - Synonym
  - Context
  - Language culture

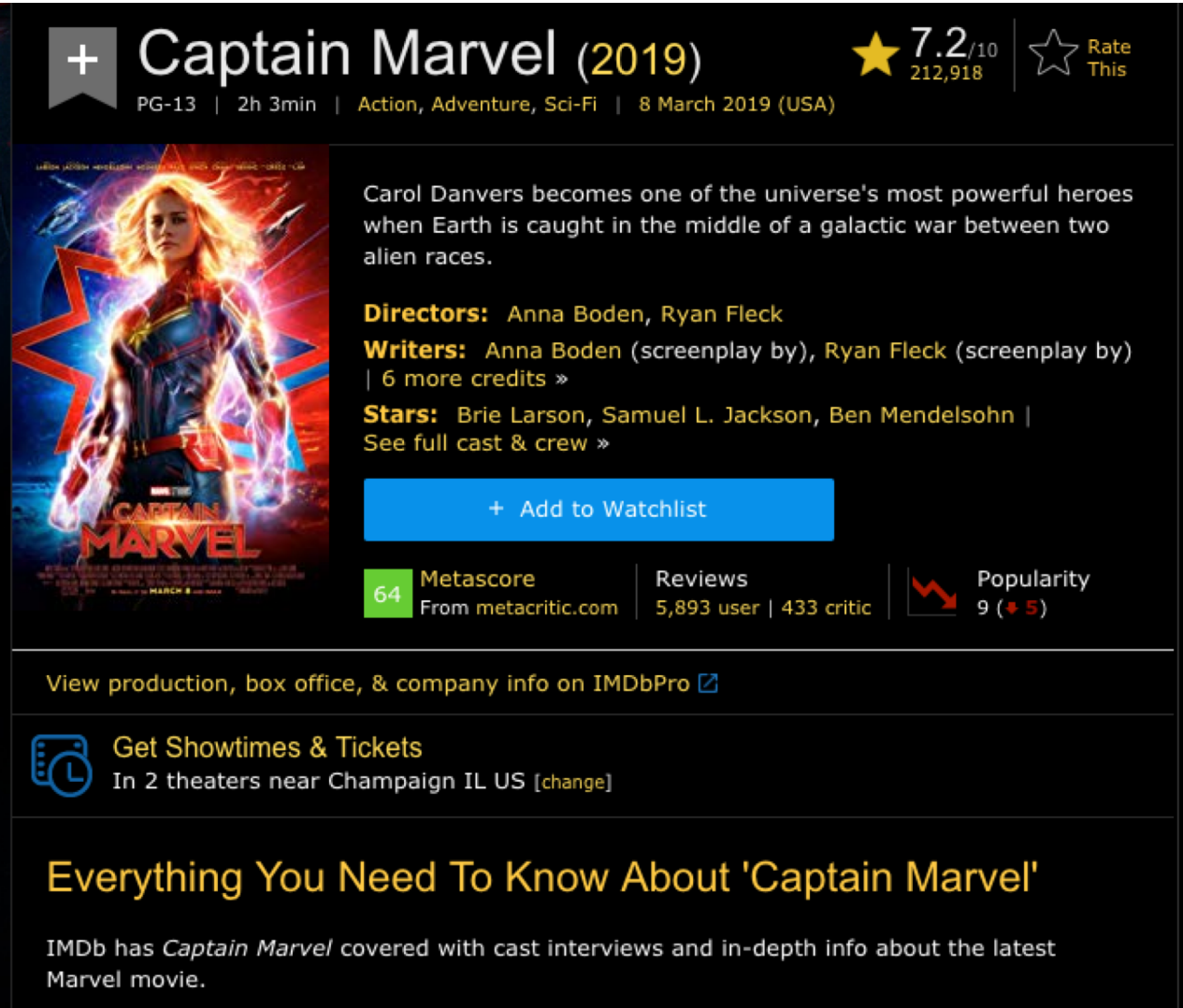
Sentence level

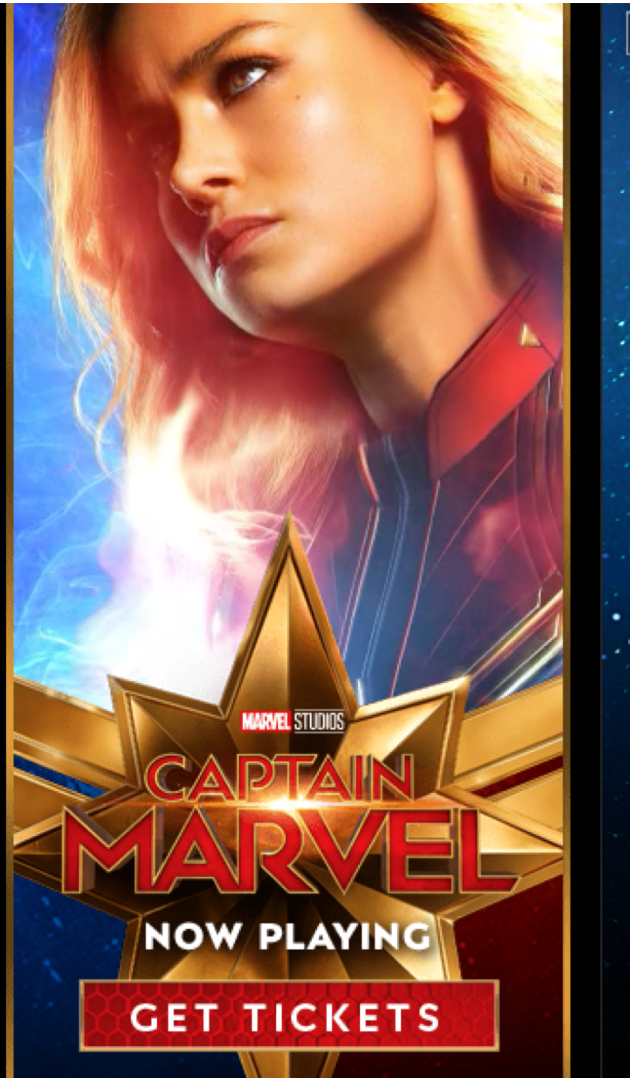
- Using words and corpus to help.
  - Bag of words; normalized sum score.
  - Pay attention of bag of words

Document level

- Hard to judge
  - Complex (too many components)
  - Complicated (hard)
- Overall rating could help somehow.
- Examples from IMDB

https://www.imdb.com/title/tt4154664/?ref_=nv_sr_1?ref_=nv_sr_1





User Reviews

User Reviews

★★★★☆☆☆☆☆☆ **What is there to marvel at?**

6 March 2019 | by EnoVarma – See all my reviews

I was left with the general feeling, that "Captain Marvel" is a major disappointment.

First of all, there is a good story here somewhere, but it's just not well told, and there's too much of it crammed within about 115 minutes. The movie is fast-paced, but never really works, because the pacing is flat and timing is off. Aesthetically, "Captain Marvel" has been processed through the same "Marvel filter" as the rest of them. Dull. The filmmakers (indie directors of the great "Half Nelson") have given absolutely no attention to the music: the score is completely forgettable as are the 90's songs.

As for the themes, the MGTOW people are going to have a field day with "Captain Marvel's" less than subtle "feminist agenda". I'm putting that in quotations, because it's more accurately described as simply female agenda. Which is absolutely fine and commendable; the problem is the infantile in-your-face way the filmmakers address the theme. The movie is set in the year 1995, and it really feels made for that year's audience. "Captain Marvel" lacks a fresh perspective. And, by the way, any fans of "Buffy, The Vampire Slayer" will immediately

datasets

- Try analyze these 20 movie reviews in your assignment 8.
  - Label.
  - Processing.
  - Output.
  - Discussion.

Last but not least…

- Feedback
- Assignment 7
- Assignment 8
- Finals

Assignment 8

- Word cloud (using tf-idf to determine font size)
- Write how you do the text processing
- Hint: you could use either binary or trinary classes.

For the rest of semesters.

- Students' feedback
- Project
  - Presentation (30min + Q&A)
  - Report (5-10 pages, academic structure)
- Final

- Have a good rest of your week.