Text Mining / Social Media Analysis

Kenn H. Kim, Ph. D.

School of Business
Clemson University

1

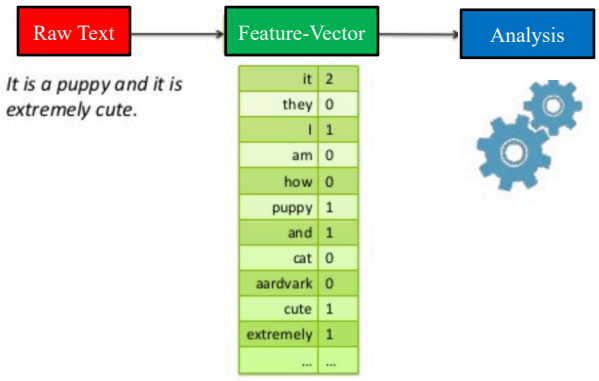
Key Concepts and Techniques

- **Key Concepts**
  - Text representation
  - Bag of words
  - Text preprocessing
- **Key Techniques**
  - Term frequency (TF)
  - Inverse Document Frequency (IDF)
  - TF-IDF
  - Sentiment Analysis
  - Topic Modeling (NMF)

2

Text Representation

- Take a set of documents (i.e. Raw Text) – each of which is a relatively free-form sequence of words – and turn it into our familiar feature-vector form (e.g. Bag of Words, Word Embedding, etc.).



3

Text Representation

- **Basic Terminology**
  - **Document**
    - A document is one piece of text, no matter how large or small. (e.g., a single sentence, a 10-page report, etc.)
  - **Token/Term**
    - A document is composed of individual tokens/terms (e.g. a word).
  - **Corpus**
    - A collection of documents is called a corpus (Latin for "body").

4

Bag of Words: Term Frequency (TF)

- Treat every document as just a collection of individual words
  - Ignore grammar, word order, sentence structure, and (usually) punctuation
  - Treat every word in a document as a potentially important keyword of the document

- **What will be the feature's value in a given document**?
  - Each document is represented by the ***term frequency*** (i.e. the word count in the document).
  - ***The importance of a term*** in a document should increase with ***the number of times that the term occurs***.

5

Bag of Words: TF Example

CLEMSON College of BUSINESS

- **An Example of Three Simple Documents (D1, D2, and D3)**
  - D1: jazz music has a swing rhythm
  - D2: swing is hard to explain
  - D3: swing rhythm is a natural rhythm

	jazz	music	has	a	swing	rhythm	is	hard	to	explain	natural
D1	1	1	1	1	1	1	0	0	0	0	0
D2	0	0	0	0	1	0	1	1	1	1	0
D3	0	0	0	1	1	2	1	0	0	0	1

6

Text Pre-processing

The following steps should be performed:

- **The case should be normalized (jazz vs. Jazz)**
  - Every term should be in lowercase.
- **Words should be stemmed (boy vs. boys)**
  - Suffixes are removed.
  - e.g., noun plurals are transformed to singular forms
- **Stop-words should be removed (a, an, the, and, …)**
  - A stop-word is a very common word in English.
  - Typical words such as *the*, *and*, *of*, and *on* are removed.

7

Text Pre-processing: Example

1) The case should be normalized.

- *Microsoft Corp and Skype Global today announced that they have entered into a definitive agreement under which Microsoft will acquire Skype, the leading Internet communications company, for $8.5 billion in cash from the investor group led by Silver Lake. The agreement has been approved by the boards of directors of both Microsoft and Skype.*

8

Text Pre-processing: Example

2) Words should be stemmed.

- *microsoft corp and skype global today announced that they have entered into a definitive agreement under which microsoft will acquire skype, the leading internet communications company, for $8.5 billion in cash from the investor group led by silver lake. the agreement has been approved by the boards of directors of both microsoft and skype.*

9

Text Pre-processing: Example

3) Stop-words should be removed.

- *microsoft corp and skype global today announce that they have enter into a definit agreement under which microsoft will acquire skype, the lead internet communic compani, for $8.5 billion in cash from the invest group lead by silver lake. the agreement has approve by the board of director of both microsoft and skype.*

10

Text Pre-processing: Example

- *microsoft corp skype global today announce enter definit agreement microsoft acquire skype lead internet communic compani billion cash invest group lead silver lake agreement approve board director microsoft skype*

Term	Count	Term	Count	Term	Count	Term	Count
skype	3	microsoft	3	agreement	2	global	1
approv	1	announc	1	acquir	1	lead	1
definit	1	lake	1	communic	1	internet	1
board	1	led	1	director	1	corp	1
compani	1	investor	1	silver	1	billion	1

11

Bag of Words: IDF

- ***Term frequency (TF)*** measures how ***prevalent*** a term is in ***a single document***.
- Now, let's think about ***a term in the entire corpus***, a collection of documents.
- When deciding the weight (i.e. importance) of a term, we may also care how **sparse** it is ***in the entire corpus (i.e. among documents)***.
- This is the **Inverse Document Frequency (IDF).**

12

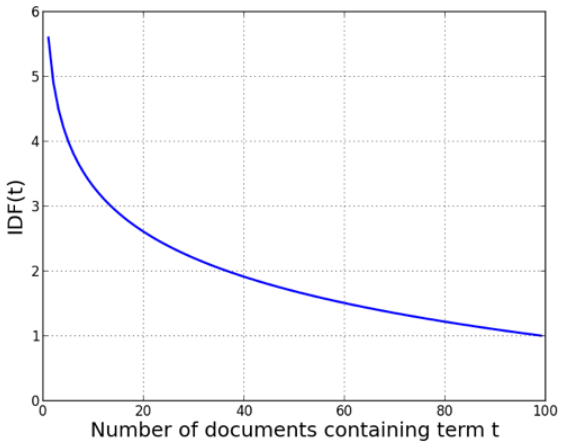
Bag of Words: IDF Example

- When deciding the weight (i.e. ***importance***) of a term, we may also care how **sparse** it is ***in the entire corpus (i.e. among documents)***. This is the **Inverse Document Frequency (IDF).**
- **An Example of Three Simple Documents (D1, D2, and D3)**
  - D1: jazz music has a unique **rhythm**
  - D2: **rhythm** is hard to explain
  - D3: **swing** is a natural **rhythm**
- **rhythm vs. swing in the entire corpus**
  - **rhythm** is "*common" (not a important term of D1, D2, or D3)*
  - **swing** is "*sparse" (a very important term of D3).*

13

Bag of Words: IDF

$$\mathrm{IDF}(t) = 1 + \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing } t}\right)$$



14

Bag of Words: TFIDF

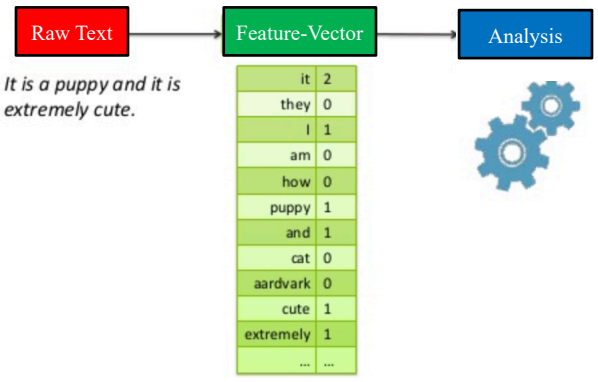
- ***Term frequency (TF)*** measures how ***prevalent*** a term is in ***a single document***.
- ***Inverse document frequency (IDF)*** measures the ***importance*** of a term ***in the entire corpus***, a collection of documents.
- **TFIDF** is a very popular representation for text. It **combines the TF and the IDF**.
  - *The TFIDF value of a term t in a given document d is thus:*

$$TFIDF(t,d) = TF(t,d) \times IDF(t)$$

15

Text Representation: Bag of Words

- **Text Representation**
  - Take a set of documents (i.e. Raw Text) – each of which is a relatively free-form sequence of words – and turn it into our familiar feature-vector form (e.g. Bag of Words).



16

Text Representation: Bag of Words

- Treat every document as just a collection of individual words
  - Ignore grammar, word order, sentence structure, and (usually) punctuation

a) ***Term frequency (TF)*** measures how ***prevalent*** a term is in ***a single document***.

b) ***Inverse document frequency (IDF)*** measures the ***importance*** of a term ***in the entire corpus***, a collection of documents.

c) **TFIDF** is a very popular representation for text. It **combines the TF and the IDF**.

- $TFIDF(t,d) = TF(t,d) \times IDF(t)$

17

Sentiment Analysis

CLEMSON College of BUSINESS

❑ Sentiment Analysis

- We also want to be able to recognize whether ***the writer's attitude*** towards a particular topic ***is positive, negative, or neutral.***

*a) It is **fun** and **easy** to do sentiment analysis!* ***(+2 positive)***

*b) I **hate** python. It makes me **frustrated**…* ***(-2 negative)***

1) Polarity

- It represents the positivity (or negativity) of a given text.

2) Subjectivity

- It represents the subjectivity (or objectivity) of a given text.
- Removing objective sentences from a document before classifying its polarity helped improve performance.

18

Sentiment Analysis: Example

CLEMSON College of BUSINESS

❑ Sentiment Analysis (Example #1)

*a) It is **fun** but **hard** to do sentiment analysis! (P:1) (N:1) (T:9)*

*b) I **love** BA but **hate** coding; it makes me **frustrated**. (P:1) (N:2) (T:10)*

Let's define the subjectivity score to be the share of positive/negative words

- subjectivity = (# of pos/neg words) / (# words)
- If there is no emotional keyword, then subj = 0
- If all words are emotional, then subj = 1

Then let's define the polarity score to be as following:

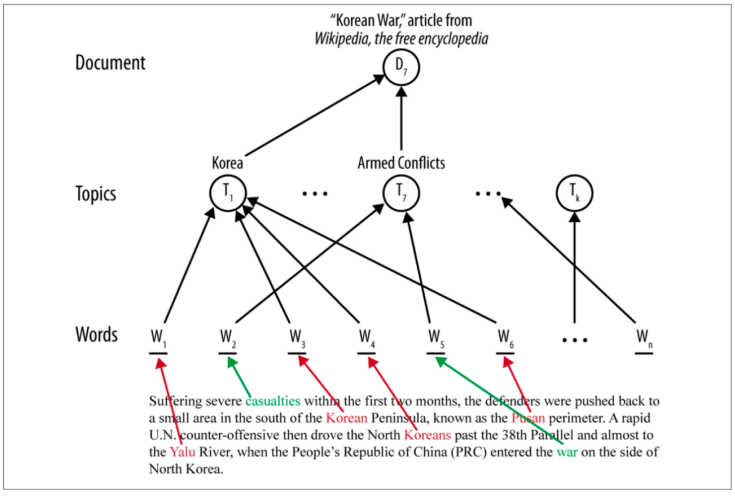
- polarity = (# pos words - #neg words) / (# words)
- If all words are postive, then pol = 1
- If all words are negative, then pol = -1
- If # pos words == # neg words, then pol = 0

19

Topic Modeling

CLEMSON College of BUSINESS

❑ Latent Information Model (i.e. Topic Modeling)

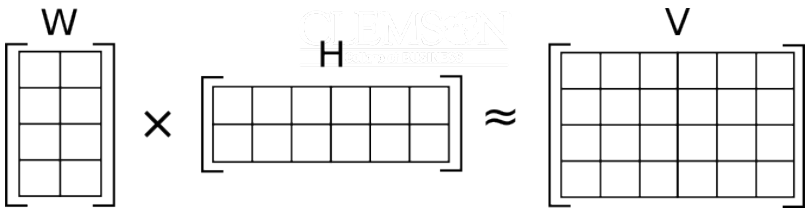


20

Topic Modeling: NMF

❑ **Latent Information Model (i.e. Topic Modeling)**

- **Non-Negative Matrix Factorization (NMF)**
  - **NMF is based on "Linear Algebra".**



21

Topic Modeling: NMF Example

❑ **Latent Information Model (i.e. Topic Modeling)**

- A ~ WH
- Tweet 1
- Tweet 2
- Tweet 3

Term-Tweet Matrix

	Word 1	Word 2	Word n
Tweet 1	1	0	2
Tweet 2	0	1	0
Tweet 3	0	1	1

Specify No Themes (k)

Features Matrix

	Word 1	Word 2	Word n
Theme 1	0.5	0	1
Theme 2	0	0.5	0

Weights Matrix

	Theme 1	Theme 2
Tweet 1	1	0
Tweet 2	0	1
Tweet 3	0	1

22

CLEMSON
College of BUSINESS

Thanks for listening!

23