Text Mining: Overview, Applications and Issues

CSE634
Data Mining
Professor Anita Wasilewska

Sources cited

[1] Text Mining: Concepts, Applications, Tools and Issues – An Overview, International Journal of Computer Applications (0975 – 8887) Volume 80 – No.4, October 2013

[2] Wasilewska, Anita. (2016). “Data Mining”. The State University of New York at Stony Brook. CSE 537 Spring 2016, “http://www3.cs.stonybrook.edu/~cse634/16ch1DMintrod.pdf “

[3] "Content Analysis of Verbatim Explanations". Ppc.sas.upenn.edu. Retrieved2015-02-23

[4] A Business Intelligence System, H.P. Luhn, IBM journal article, 1958

[5]“Getting started in text mining”, Cohen, K. Bretonnel; Hunter, Lawrence (2008). PLoS Computational Biology 4 e20. doi: 10. 1371/journal.pcbi.0040020.

[6] ”Text Analytics”, Medallia, Retrieved 2015-02-23

[7] “Unstructured data and 80 percent rule”, Seth Grimes; Breakthrough analysis, Retrieved 2015-02-23

[8] “Text mining tools techniques and applications”, Nathan Treloar, Aquaquest Inc.

[9] http://kaylinwalker.com/text-mining-south-park/

[10] Common Text Mining Workflow by Ricky Ho

[11] A survey of different text mining techniques by Varsha C. Pande and Dr A.S.Khandelwal

[12] Overview and Semantic Issues of Text Mining, Anna Stavrianou, Periklis Andritsos, Nicolas Nicoloyannis, SIGMOD Record, September 2007 (Vol. 36, No. 3)

Overview

- Early history
- Applications
- Introduction to text mining
- Need for text mining
- Challenges in text mining
- Text mining process
- Areas of text mining
- Case study: Text Mining South Park

Brief Early History

- Manual text mining approaches first surfaced in mid 1980's[3]
- The challenge of exploiting the large proportion of enterprise information that originates in "unstructured'' form was first recognized in IBM Journal article by H.P. Luhn[4]
- As BI emerged in the '80s and '90s as a software category, the emphasis was on numerical data stored in relational databases.
- Technological advances have enabled the field to advance during the past decade

[3] "Content Analysis of Verbatim Explanations". Ppc.sas.upenn.edu. Retrieved2015-02-23
[4] A Business Intelligence System, H.P. Luhn, IBM journal article, 1958

Applications

- The technology is now broadly applied for a wide variety of government, research and business needs.
- Application categories include:
  - Security applications - monitoring online text sources
  - Biomedical applications - knowledge based search engine for biomedical texts.
  - Marketing applications - analytical customer relationship management, improve predictive analysis models, etc.
  - Sentiment analysis - analysis of data for predicting desired results
  - Software applications – used by major firms to automate analysis
  - Academic, online media, digital humanities, etc.

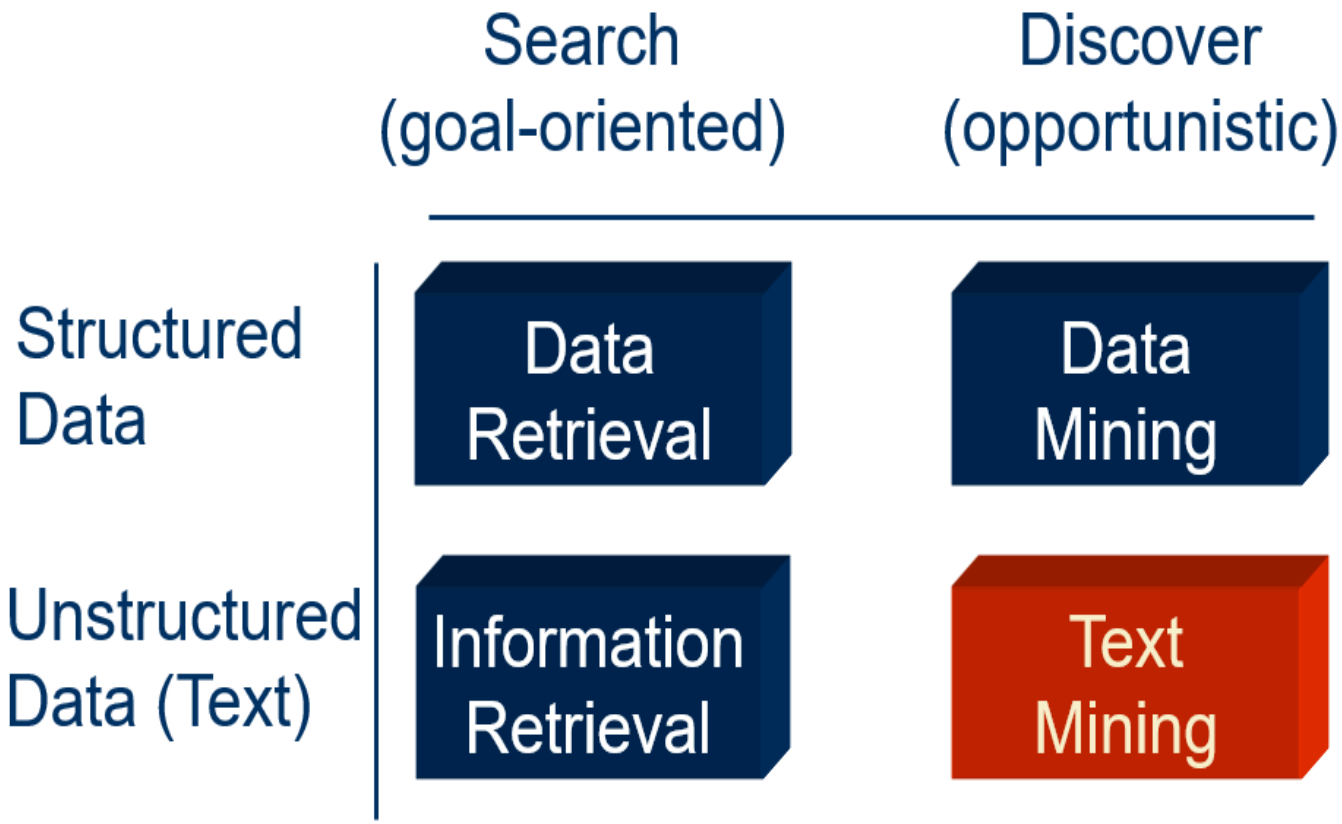
"Getting started in text mining", Cohen, K. Bretonnel; Hunter, Lawrence (2008). PLoS Computational Biology 4 e20. doi: 10. 1371/journal.pcbi.0040020.
"Text Analytics", Medallia, Retrieved 2015-02-23

What is Text Mining?

- It is also referred as text data mining and is roughly equivalent to text analytics.
- Process of extracting interesting and non-trivial knowledge from unstructured text.
- Text analysis also involves following:
  - information retrieval
  - lexical analysis to study word frequency distribution
  - pattern recognition
  - predictive analytics
  - visualization, etc.
- Overall goal is to turn text into data for analysis via application of Natural Language Processing and analytical methods.

"Unstructured data and 80 percent rule", Seth Grimes; Breakthrough analysis, Retrieved 2015-02-23

Search vs Discover



"Text mining tools techniques and applications", Nathan Treloar, Aquaquest Inc.

Data Retrieval

- It find records within a structured database.
- **Database type:** Structured
- **Search mode:** Goal-dirven
- **Example Information need:** “Find a restaurant that serves vegetarian food”

Information Retrieval

- It finds a relevant information in unstructured information source.
- **Database type:** Unstructured
- **Search mode:** Goal - driven
- **Example Information need:** “Find a restaurant that serves vegetarian food”

Data Mining

- It discovers new knowledge through analysis of data.
- **Database type:** Structured
- **Search mode:** Opportunistic
- **Example Information need:** “Find the trend over number of visits to a vegetarian restaurant”

Text Mining

- It discovers new knowledge through analysis of text.
- **Database type:** Unstructured
- **Search mode:** Opportunistic
- **Example Information need:** “Find the types of food poisoning most often associated with junk food”

Need for Text Mining

- Vast amounts of new information and data are generated everyday through economic, academic and social activities, much with significant potential economic and societal value. Techniques such as text and data mining and analytics are required to exploit this potential.
- Approximately **90%** of the world's data is held in unstructured formats (source: Oracle Corporation)

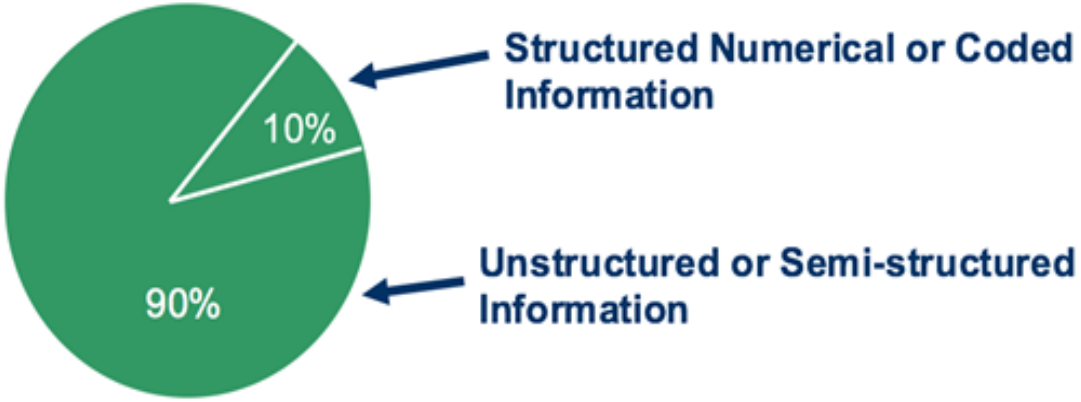


Image source: 2002, AvaQuest Inc.

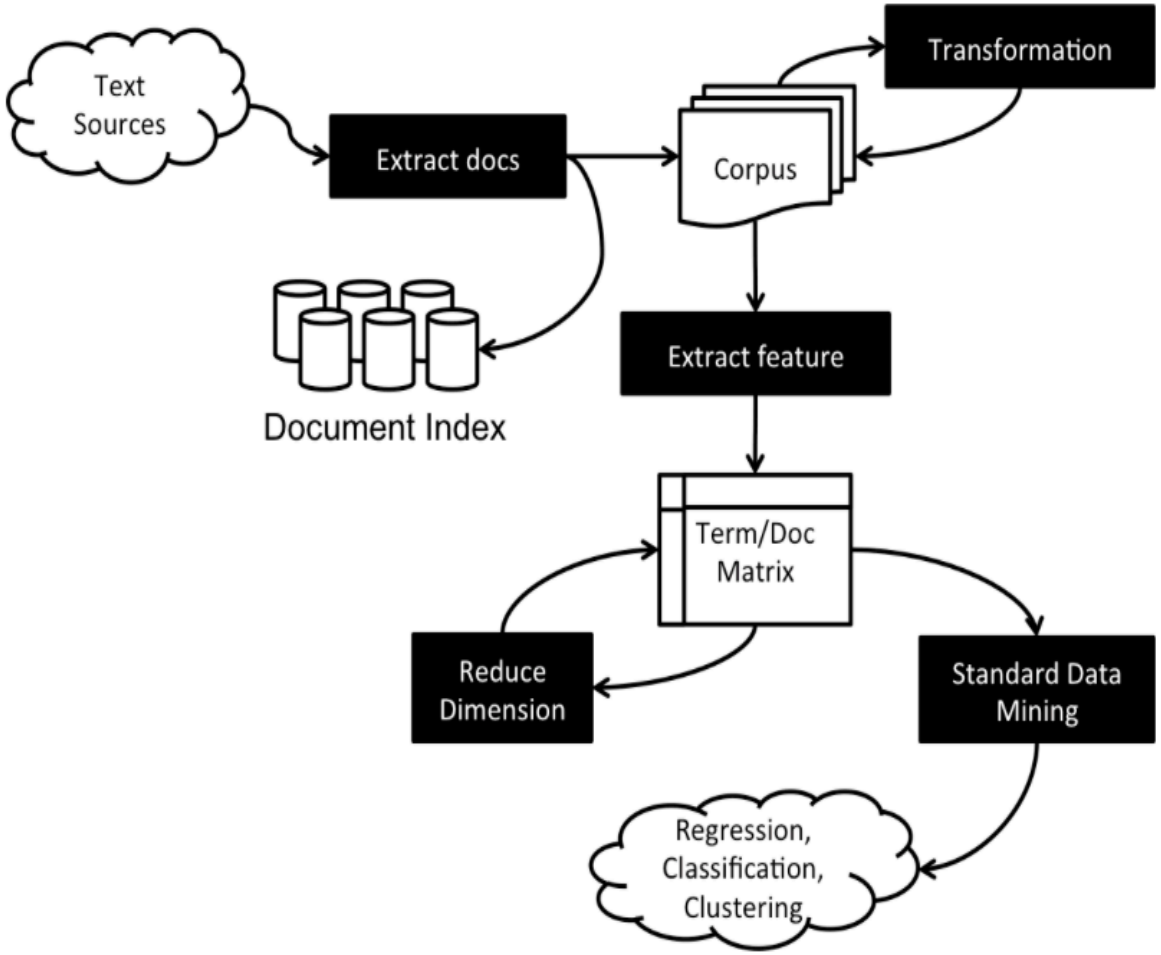
Challenges in Text Mining

- Large datasets
- Noisy data
- Word Ambiguity and Context Sensitivity
  - Apple (the company) or apple (the fruit)
- Context Sensitivity
  - automobile = car = vehicle = Toyota
- Complex and subtle relationship between concepts in text
  - Eg: “AOL merges with Time-Warner” “Time-Warner is bought by AOL”
- Multilingual

Text Mining Process

- Extract Documents
- Text Transformation
- Feature Extraction
- Reduce Dimensions
- Apply standard Data Mining
- Interpretation/Evaluation

Common Text Mining Workflow by Ricky Ho



https://dzone.com/articles/common-text-mining-workflow

Extract Documents

- In this phase, we are extracting text document from various types of external sources into a text index (for subsequent search) as well as a text corpus (for text mining).
- Document source can be a public web site, an internal file system, etc.

Common Text Mining Workflow by Ricky Ho

How to?

- Perform a google search or crawl a predefined list of web sites, then download the web page from the list of URL, parse the DOM to extract text data from its sub-elements, and eventually creating one or multiple documents, store them into the text index as well as text Corpus.
- Invoke the Twitter API to search for tweets (or monitor a particular topic stream of tweets), store them into the text index and text Corpus.
- If the text is in a different language, we may also invoke some machine translation service (e.g. Google translate) to convert the language to English.

Common Text Mining Workflow by Ricky Ho

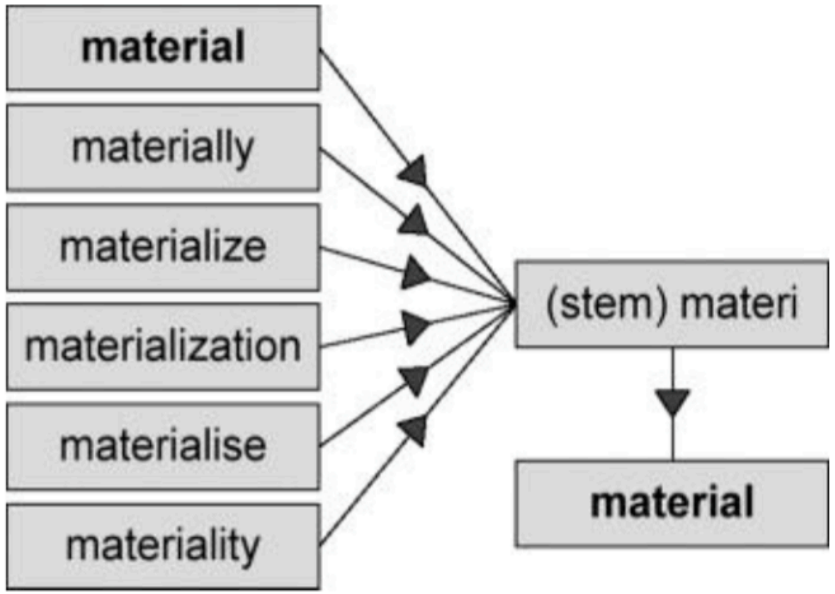
Text Preprocessing and Transformation

- Tokenization: Text documents contain a collection of statements. This step segments the whole text into words by removing blank spaces, commas etc.

- Stop word Removal: This step involves removing of HTML, XML tags from web pages. Then the process of removal of stop words such as ‘a’, ‘is’, ‘of’, etc is performed.

Common Text Mining Workflow by Ricky Ho

- Stemming: Stemming refers to the process of identifying the root of a certain word.



Common Text Mining Workflow by Ricky Ho

- To extract information about some entities mentioned in the document, we need to conduct sentence segmentation, paragraph segmentation in order to provide some local context from which we can analyze the entity with respect to its relationship with other entities.

- Attach Part-Of-Speech tagging, or Entity tagging (person, place, company) to each word.

- Apply standard text processing such as lower case, removing punctuation, removing numbers, removing stop word, stemming.

- Optionally, normalize the words to its synonyms using Wordnet or domain specific dictionary.

Common Text Mining Workflow by Ricky Ho

Feature Selection

- It is also known as variable selection.
- It is the process of selecting a subset of important features for use in model creation.
- This phase mainly performs removing features which are redundant or irrelevant.
- For text mining, the "bag-of-words model" is commonly used as the feature set.
- After this phase, the Corpus will turn into a large and sparse document term matrix.

Common Text Mining Workflow by Ricky Ho

Reduce Dimensions

- **Why?**
- For efficiency reason, we want to reduce the memory footprint for storing the corpus
- We want to transform the vector from the "term" space to a "topic" space, which allows document of similar topics to situate close by each other even they use different terms. (e.g. document using the word "pet" and "cat" are map to the same topic based on their co-occurrence)

How?

- SVD (Singular Value Decomposition) is a common matrix factorization technique to convert a "term" vector into a "concept" vector. SVD can be used to factor a large sparse matrix of M by N into the multiplication of three smaller dense matrix M*K, K*K, K*N.

- Another popular technique called topic modeling is also commonly used to transform the document into a smaller set of topic dimensions.

Common Text Mining Workflow by Ricky Ho

Text Mining Techniques

- Document Clustering
  - Text Categorization
  - Text Clustering
  - Sentiment Analysis

A survey of different text mining techniques by Varsha C. Pande and Dr A.S.Khandelwal

Document Clustering

- It is the process of application of cluster analysis to textual data
- Two algorithms:
  - Hierarchical- Agglomerative and Decisive
  - K-means and its variant
- Applications:
  - Document Organization and browsing
  - Corpus Summarization

A survey of different text mining techniques by Varsha C. Pande and Dr A.S.Khandelwal

Text Categorization

- It aims at assigning documents to one or more classes or categories.
- Content-based classification: the weight given to particular subjects in a document determines the class to which the document is assigned.
- Applications:
  - Sentiment Analysis
  - Language identification of text

A survey of different text mining techniques by Varsha C. Pande and Dr A.S.Khandelwal

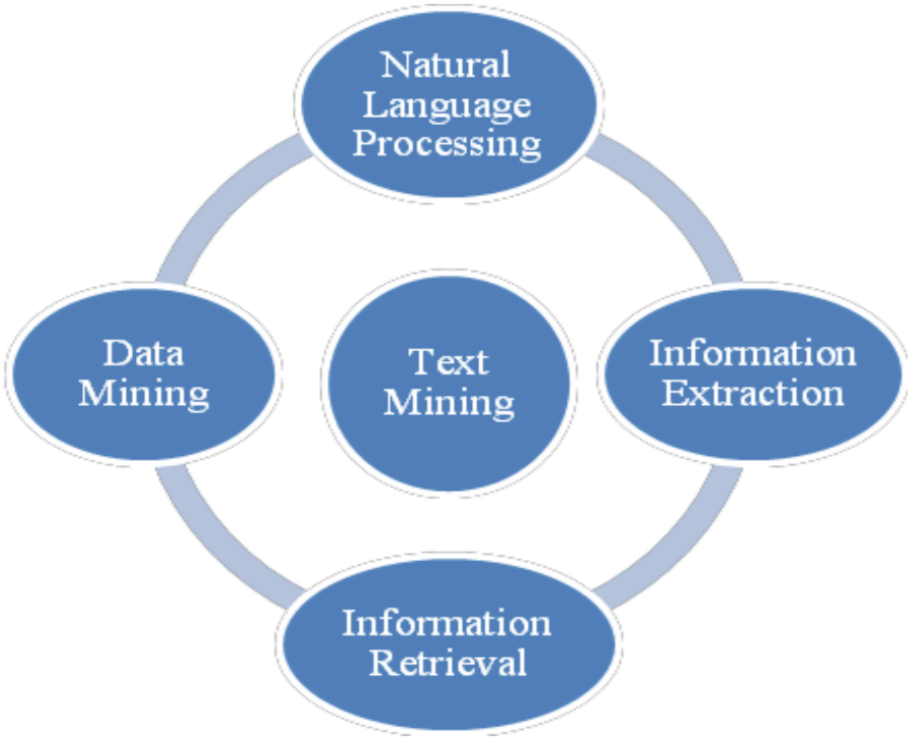
Interpretation/ Evaluation

- Analyzing the results
- Check accuracy
- Repeat the algorithm with refined data

Common Text Mining Workflow by Ricky Ho

Areas of Text Mining

- Information Extraction
- Information Retrieval
- Natural Language Processing
- Data Mining



Text Mining: Concepts, Applications, Tools and Issues – An Overview, International Journal of Computer Applications (0975 – 8887) Volume 80 – No.4, October 2013

Information Extraction

- It is the process of automatically extracting structured information from unstructured and/or semi structured text documents.
- Pattern matching is the output of the IE process.
- Functions performed by IE Systems are:
  - Term Analysis
  - Named Entity Recognition
  - Fact Extraction

Text Mining: Concepts, Applications, Tools and Issues – An Overview, International Journal of Computer Applications (0975 – 8887) Volume 80 – No.4, October 2013

Information Retrieval

- It is defined as the methods used for representation, storage and accessing of information items where the information handled is mostly in the form of textual documents, newspapers and books which are retrieved from databases according to the user request or queries.
- An IR system allows us to narrow down the set of documents that are relevant to a specific problem. The most well known IR systems are search engines such as Google.

Text Mining: Concepts, Applications, Tools and Issues – An Overview, International Journal of Computer Applications (0975 – 8887) Volume 80 – No.4, October 2013

Natural Language Processing

- The role of NLP in text mining is to provide the Information Extraction Phase with linguistic data that they need to perform their task.
- Often it is annotating documents with information like sentence boundaries, part-of-speech tagging, parsing which can then be read by the information extraction tools.

Text Mining: Concepts, Applications, Tools and Issues – An Overview, International Journal of Computer Applications (0975 – 8887) Volume 80 – No.4, October 2013

Case Study : Text Mining South Park

- South Park follows four fourth grade boys (Stan, Kyle, Cartman and Kenny) and an extensive ensemble cast of recurring characters.
- This analysis reviews their speech to determine which words and phrases are distinct for each character.

http://kaylinwalker.com/text-mining-south-park/

How it is done?

- The programming language R and packages tm, RWeka and stringr were used to scrape South Park episode transcripts from the internet.
- Preprocessing and transformation
- Calculate the log likelihood for each character pair, and rank them to create a list of most characteristic words/phrases for each character.
- The results were visualized using ggplot2, wordcloud and RColorBrewer.

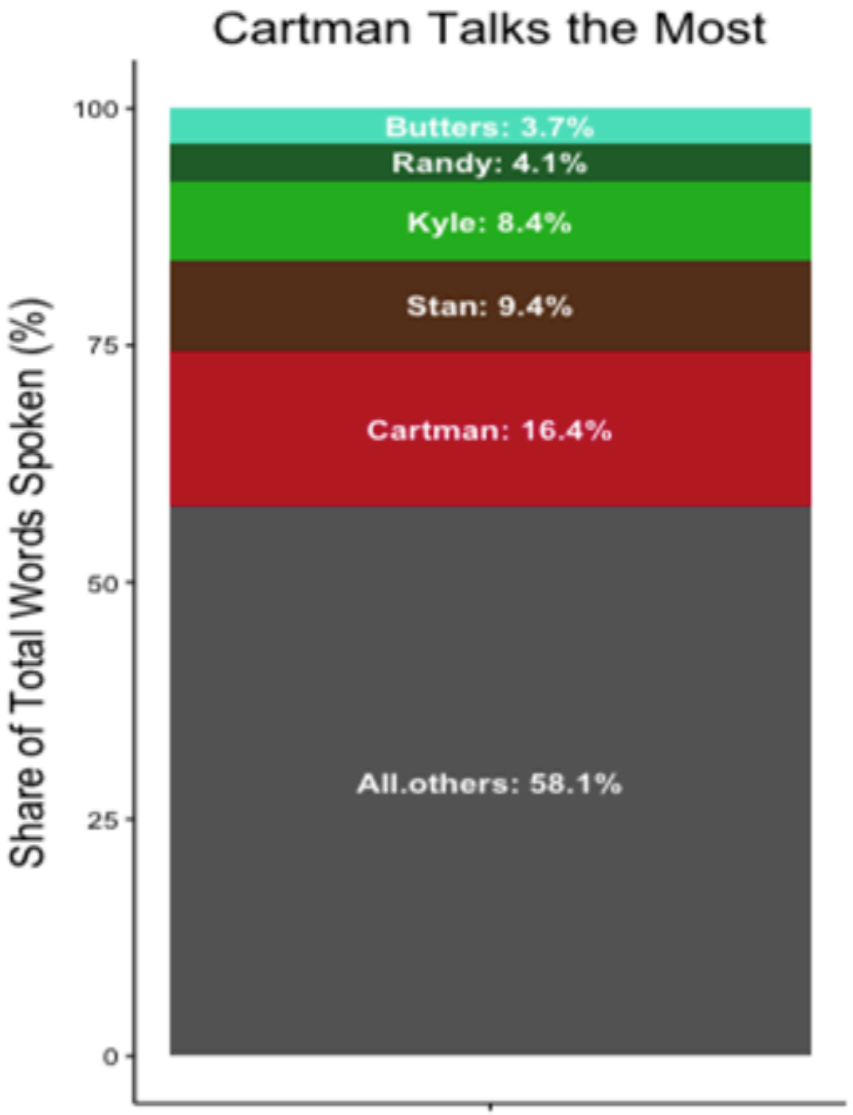
Word cloud after initial steps

Number of words spoke per character

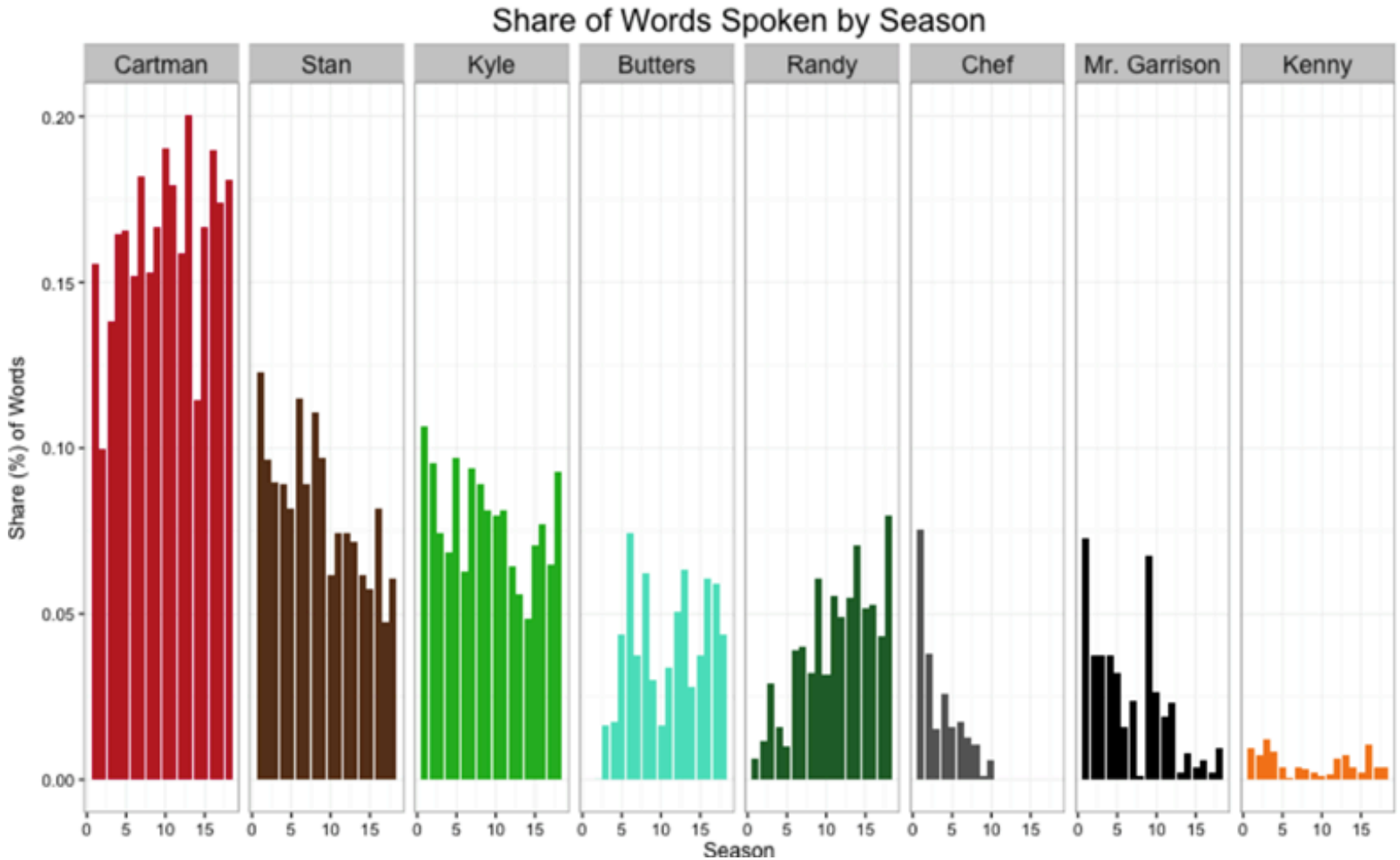
Number of Words by Character

speaker	words	speaker	words	speaker	words
cartman	61110	jimmy	3738	narrator	1737
stan	34762	gerald	3285	principal.victoria	1732
kyle	31277	jimbo	3157	jesus	1714
randy	14994	announcer	2900	mayor	1603
butters	13690	wendy	2893	craig	1412
mr..garrison	9436	sheila	2794	reporter	1400
chef	5493	liane	2477	satan	1291
mr..mackey	4829	stephen	2245	linda	1285
sharon	4284	kenny	2112	all.others	152172

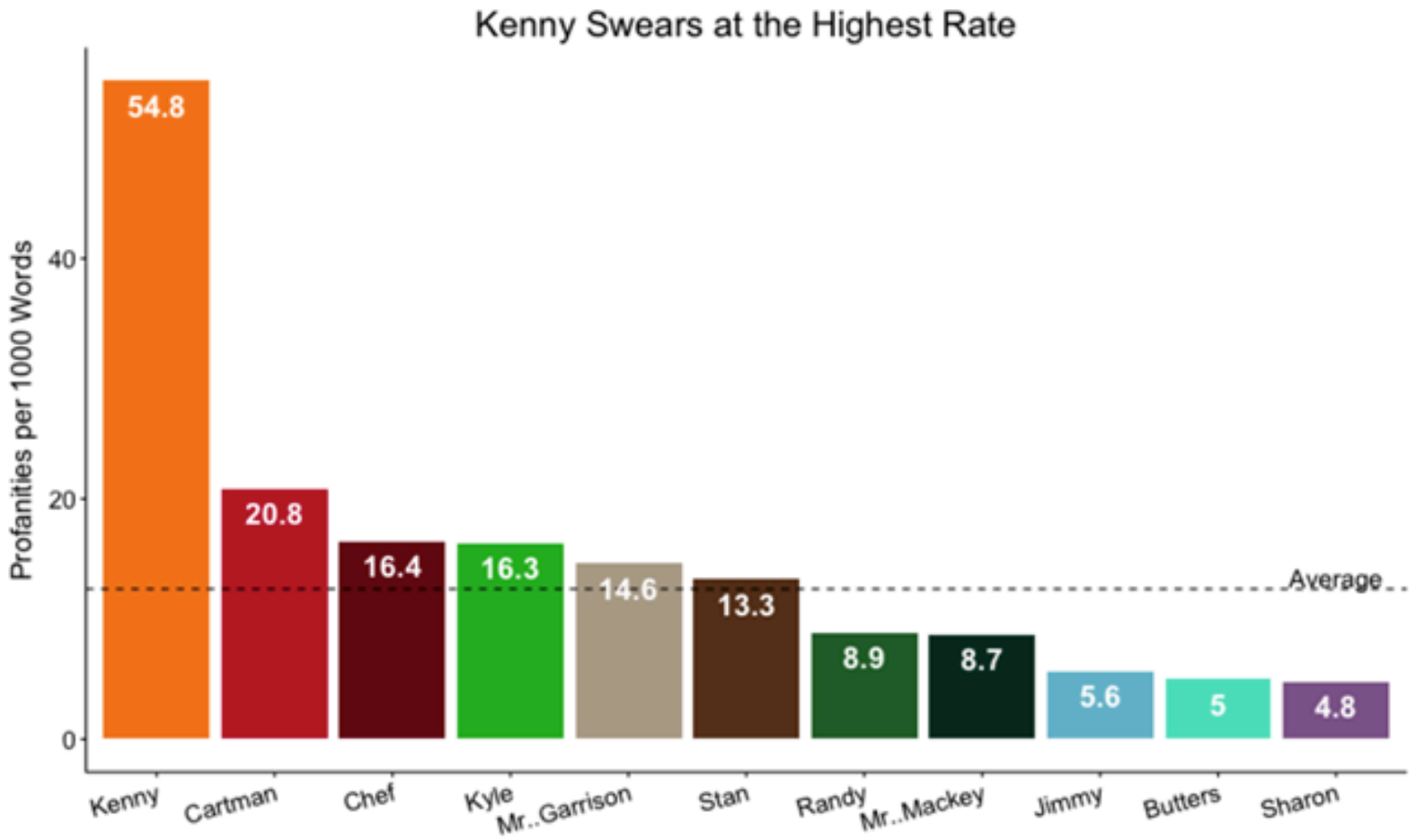
Analysis of which character talks the most



Analysis of share of words spoken across seasons



Analysis of swear words spoken by characters



Log-Likelihood

$$2 \sum O_i * ln(\frac{O_i}{E_i})$$

Which can be computed from the contingency table below as

$$(2 * ((a * log(\frac{a}{E1}) + (b * log(\frac{b}{E2}))); E1 = (a + c) * \frac{a + b}{N}; E2 = (b + d) * \frac{a + b}{N}$$

Basic Framework

Group	Corpus.One	Corpus.Two	Total
Word	a	b	a+b
Not Word	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

An Example with Log Likelihood 101.7

Group	Cartmans.Text	Remaining.Text	Total
‘hippies’	36	5	41
Not ‘hippies’	28170	144058	172228
Total	28206	144063	172269

Computed:

E1 = 28206*(41/172269) = 6.71*

E2 = 144063(41/172269) = 34.28

LL = 2*(36*log(36/6.71) + 5*log(5/34.28)) = 101.7

Analysis of characteristic words per person