Text Processing

Analysis of Summer Training Comments

Coleman Strickland, Capt USMC

https://www.rw3.com/retail-grocers-need-know-natural-language-processing/

Project Goal:

- Analyze feedback without bias
- Identify trends/common themes
- Provide a framework for future analysis

Agenda

- What is Leatherneck?
- Data Munging
- Text Mining Techniques
  - Sentiment Analysis
  - Word Groupings (n-grams)
  - Similarity Comparison
- Classifying Comments/Responses
- Real World Applications
- Common Issues/Concerns
- Easy to Use Text Mining Packages in R

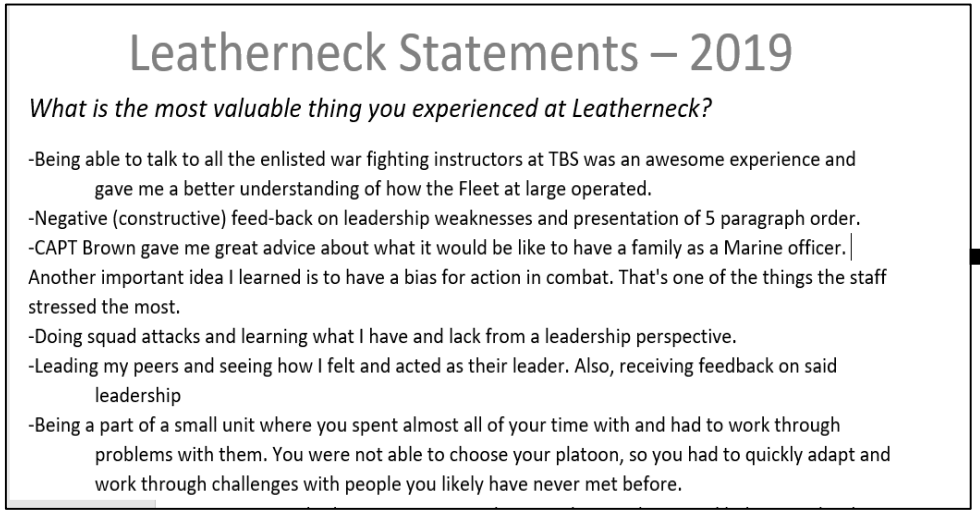
What is Leatherneck?

- Summer training event in which midshipmen are evaluated for Marine Corps Selection (directed primarily at 1/C).
- Physically demanding exercise where leadership potential is assessed and evaluated.
- Goal is to determine who has the potential to become a Marine Corps officer.

http://usnatrident.blogspot.com/2015/06/usnas-leatherneck-2015-marines-making.html

Data Munging

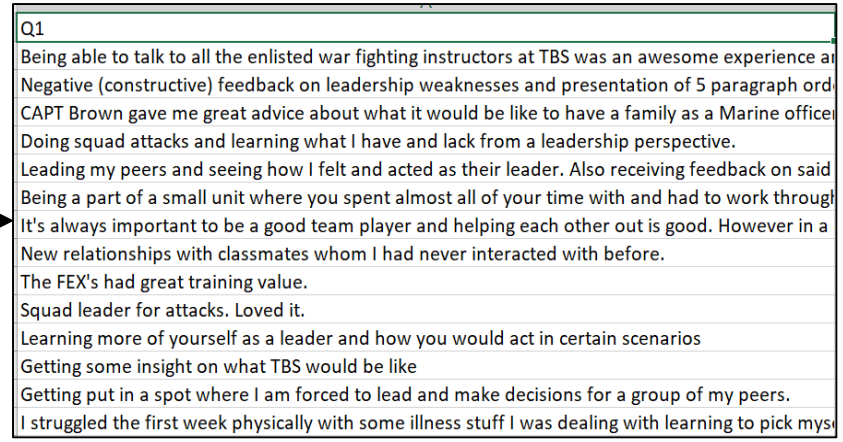
- 7 questions
- 366 comments/responses (sentences, short statements, bullets, etc.)
- 18 page Word Doc
- End State: data formatted as a .csv file (rows: comments, columns: questions)



Leatherneck Statements – 2019

What is the most valuable thing you experienced at Leatherneck?

-Being able to talk to all the enlisted war fighting instructors at TBS was an awesome experience and gave me a better understanding of how the Fleet at large operated.
-Negative (constructive) feed-back on leadership weaknesses and presentation of 5 paragraph order.
-CAPT Brown gave me great advice about what it would be like to have a family as a Marine officer.
Another important idea I learned is to have a bias for action in combat. That's one of the things the staff stressed the most.
-Doing squad attacks and learning what I have and lack from a leadership perspective.
-Leading my peers and seeing how I felt and acted as their leader. Also, receiving feedback on said leadership
-Being a part of a small unit where you spent almost all of your time with and had to work through problems with them. You were not able to choose your platoon, so you had to quickly adapt and work through challenges with people you likely have never met before.



Q1
Being able to talk to all the enlisted war fighting instructors at TBS was an awesome experience a
Negative (constructive) feedback on leadership weaknesses and presentation of 5 paragraph ord
CAPT Brown gave me great advice about what it would be like to have a family as a Marine office
Doing squad attacks and learning what I have and lack from a leadership perspective.
Leading my peers and seeing how I felt and acted as their leader. Also receiving feedback on said
Being a part of a small unit where you spent almost all of your time with and had to work through
It's always important to be a good team player and helping each other out is good. However in a
New relationships with classmates whom I had never interacted with before.
The FEX's had great training value.
Squad leader for attacks. Loved it.
Learning more of yourself as a leader and how you would act in certain scenarios
Getting some insight on what TBS would be like
Getting put in a spot where I am forced to lead and make decisions for a group of my peers.
I struggled the first week physically with some illness stuff I was dealing with learning to pick mys

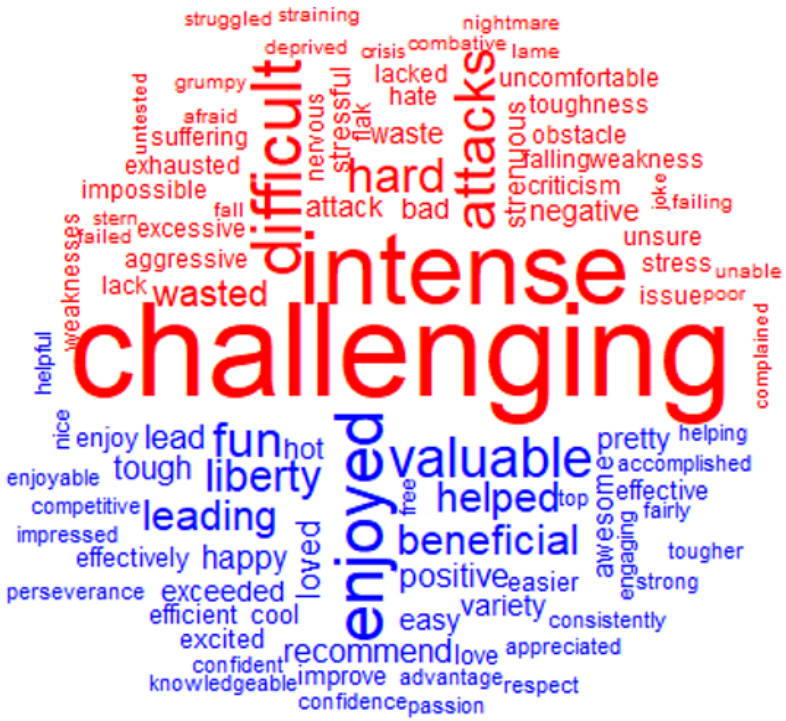
Preprocessing = Word Doc

Post processing= data frame (.csv)

Sentiment Analysis

Opinion Mining

- The use of text mining techniques to determine the attitude of the speaker.*
- Sentiment Lexicon
  - AFINN: scores individual words between -5 to 5
  - Other lexicons:
    - bing = binary score (positive and negative)
    - nrc = categorical score (positive, negative, anger, joy, trust, etc.)
- Combining sentiment scoring with word counts produces an overarching attitude toward a particular event.

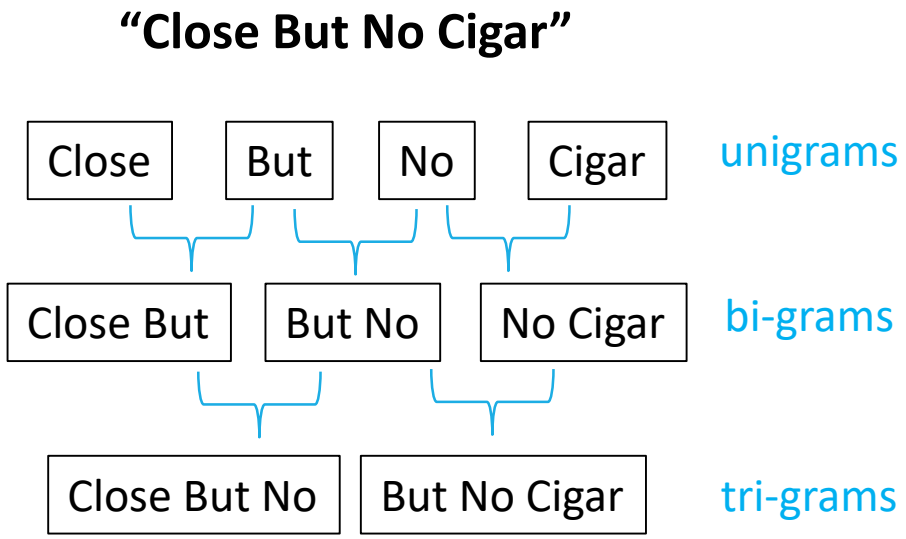


*Huddleston, Samuel PhD. Lecture: TextMining. NPS, 2017.

Word Groupings

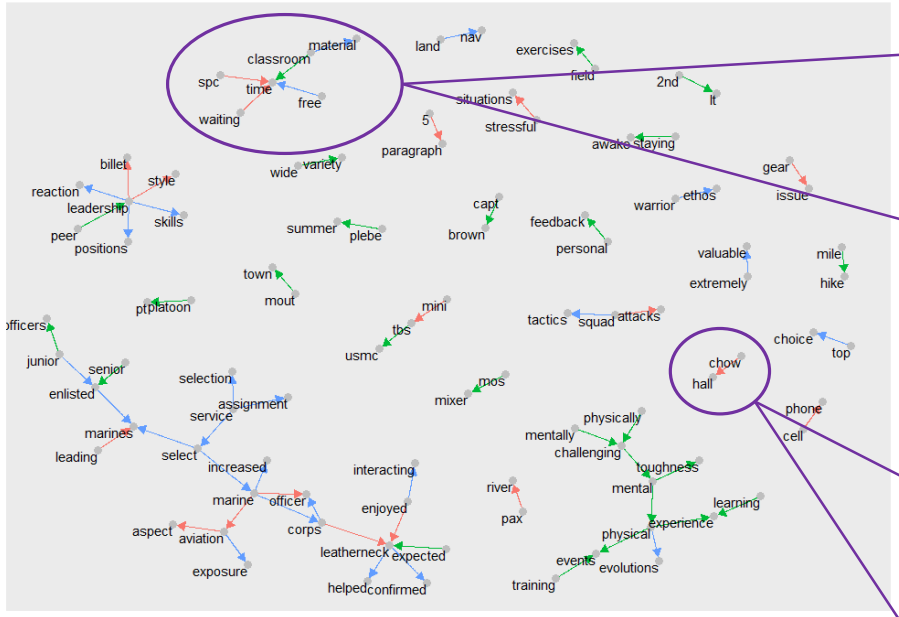
N-grams

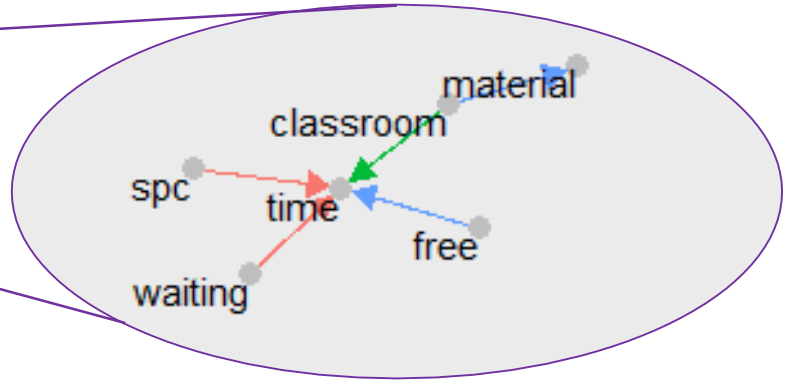
- N-grams are groupings of words/characters
- N-grams are useful in relating terms to each other (i.e. chow hall)
- As the number of N-grams grow, the amount of data stored grows

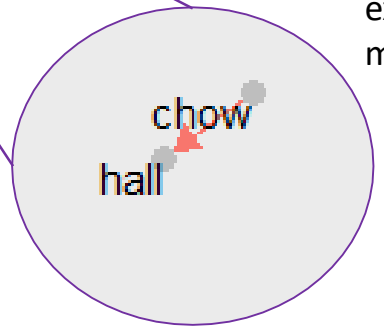


Word Groupings

Bi-gram Network with Sentiment Scoring







- Nodes = words
- Arcs = direction of use
- Colors = sentiment
  - Blue: positive
  - Green: neutral
  - Red: negative

"More coordination with the chow hall to expedite meal times and make sure there are meals everyday."

"I don't know how it would be done other than by building a new building but the chow hall failed to provide everyone with enough food in a short amount of time because so much time was wasted there."

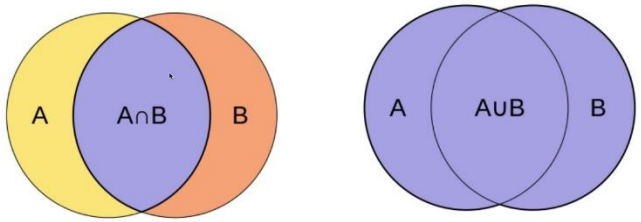
Similarity Comparison

Common Approach

- Basic Idea: two responses are similar if they contain the same terms
  - More matches = more similar
- Multiple ways to measure similarity:
  - Number of similar terms (overlap between the responses)
  - Frequency of term appearance
  - Unique terms in each response

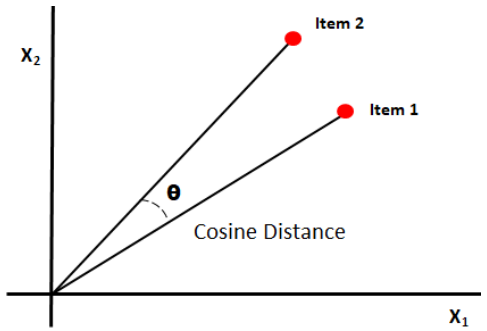
- Two Common methods:
  - Jaccard Similarity: intersection divided by union

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$



  - Cosine Similarity:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$



Similarity Comparison

Data Prep

- Prepping the text is critical to obtaining “terms” that we can use for similarity scoring.
- Tokenize → Normalize → Remove Stop Words → Stem → *N-grams?*
- Example:

Tokenize	**“Leading my peers and seeing how I felt and acted as their leader. Also receiving feedback on said leadership”**
Normalize	**“leading my peers and seeing how i felt and acted as their leader also receiving feedback on said leadership”**
Remove Stop Words	**“leading peers seeing felt acted leader also receiving feedback said leadership”**
Stem	**“lead peer see felt act leader also receiv feedback said leadership”**

Similarity Comparison

Find Common Themes

- General Approach:
  1. Prep text
  2. Compare an individual comment to all other comments inside a given question
  3. If the comparison meets a specified threshold, set the similar comment aside
  4. If three or more comments have been found, assign a group theme
  5. Repeat the process until all comments have been analyzed

- 11 common themes identified:
  - Leading Peers and Receiving Feedback (Q1)
  - TBS Exposure (Q1)
  - Aviation Exposure (Q2)
  - Liberty (Q2)
  - FEX Length (Q2)
  - Leatherneck's Impact on Service Selection (Q3)
  - Evolving Timelines (Q4)
  - 7-Mile Hike/MOUT Intensity (Q5)
  - Leatherneck was Fun (Q6)
  - Leatherneck was a Challenge (Q6)
  - Marine Corps Exposure (Q6)

7-Mile Hike/MOUT Intensity	I was expecting the 7 mile hike to be more intense but that was relatively easy. MOUT town was a bit much.
	I think the 7.8 mile hike should be more intense i.e. faster pace or more weight.
	I think that MOUT town could have been more intense. There were a few times during that evolution where it felt like more of a mess than an operation.
	I thought the 8 mile ruck was too easy.

Classifying Comments/Responses

Using Machine Learning for Classification

- Goal: Can we use the output from our similarity comparison algorithm to classify responses across questions?

- Adaptive Boosting: utilize classification trees as classifiers and use the weighted sum of the "weak learners" for final classification.
  - https://towardsdatascience.com/boosting-algorithm-adaboost-b6737a9ee60c

- Training data:
  - Create a document-term-matrix (DTM) of the classified responses (sparse matrix: 366x1160)
    - rows = documents
    - columns = all words across all documents
  - Weight the DTM using Term Frequency – Inverse Document Frequency (TF-IDF), goal is to determine the importance of each term in the document
    - $TF = \frac{\#\ of\ times\ term\ t\ appears\ in\ an\ observation}{\#\ of\ words\ in\ the\ observation}$
    - $IDF = \log \frac{\#\ of\ total\ observations}{\#\ of\ observations\ containing\ t}$
    - $TFIDF = TF * IDF$
  - Use the columns of our TF-IDF matrix as predictors in our model

Classifying Comments/Responses

Using Machine Learning to Predict Themes = TF-IDF Example

- TFIDF Example:

$$\mathrm{tf}(\text{"this"}, d_1) = \frac{1}{5} = 0.2$$
$$\mathrm{tf}(\text{"this"}, d_2) = \frac{1}{7} \approx 0.14$$
$$\mathrm{idf}(\text{"this"}, D) = \log\left(\frac{2}{2}\right) = 0$$
$$\mathrm{tfidf}(\text{"this"}, d_1, D) = 0.2 \times 0 = 0$$
$$\mathrm{tfidf}(\text{"this"}, d_2, D) = 0.14 \times 0 = 0$$

$$\mathrm{tf}(\text{"example"}, d_1) = \frac{0}{5} = 0$$
$$\mathrm{tf}(\text{"example"}, d_2) = \frac{3}{7} \approx 0.429$$
$$\mathrm{idf}(\text{"example"}, D) = \log\left(\frac{2}{1}\right) = 0.301$$
$$\mathrm{tfidf}(\text{"example"}, d_1, D) = \mathrm{tf}(\text{"example"}, d_1) \times \mathrm{idf}(\text{"example"}, D) = 0 \times 0.301 = 0$$
$$\mathrm{tfidf}(\text{"example"}, d_2, D) = \mathrm{tf}(\text{"example"}, d_2) \times \mathrm{idf}(\text{"example"}, D) = 0.429 \times 0.301 \approx 0.129$$

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

https://en.wikipedia.org/wiki/Tf%E2%80%93idf

Classifying Comments/Responses

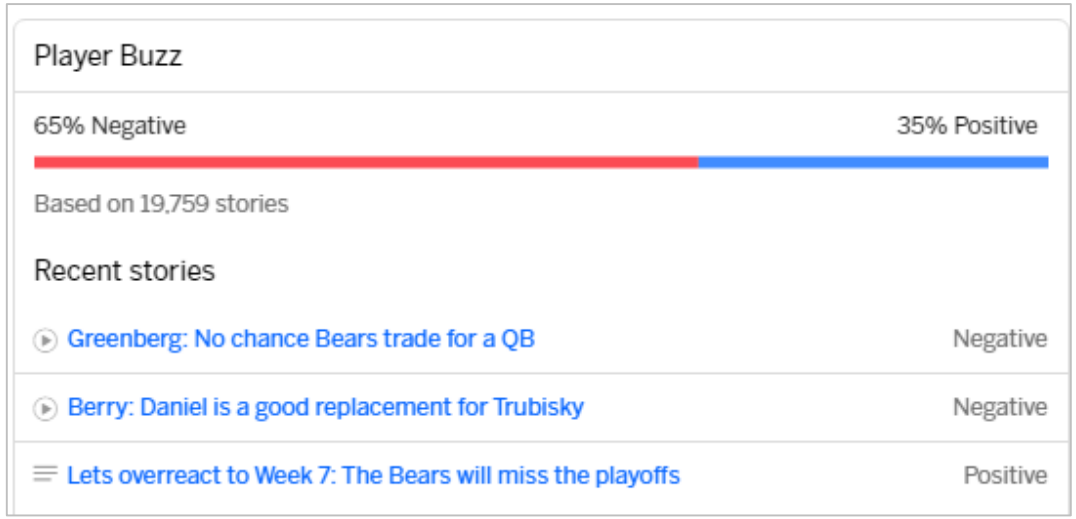
Using Machine Learning for Classification = Results

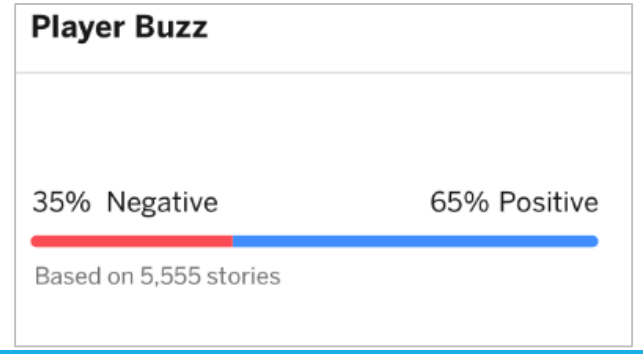
Comments	Question	Predicted Theme
I expected to be tested physically and to spend a lot of time rucking to different events. Leatherneck did match my expectations. I have heard stories about TBS from my friends and I feel like Leatherneck was a good representation of TBS.	Q6	TBS Exposure (Q1)
I garnered enough exposure to the Marine Corps to decide I belonged elsewhere.	Q1	LNK's Impact on Service Selection (Q3)
Development in teamwork skills and helping me secure my position on choosing Marine Corps first.	Q1	LNK's Impact on Service Selection (Q3)
I definitely want to service select Marine Corps	Q1	LNK's Impact on Service Selection (Q3)
More aviation stuff one day is not enough when nearly 100 mids will be Marine pilots	Q7	Aviation Exposure (Q2)
Dealing with Midshipmen that decided after a bit that they did not care very much as well as dealing with the long wait times.	Q4	Evolving Timelines (Q4)
It increased because I could see myself doing it as a career and it challenged me and caused growth	Q3	LNK was a challenge (Q6)
[The SPC] gave me great advice about what it would be like to have a family as a Marine officer. Another important idea I learned is to have a bias for action in combat. That's one of the things the staff stressed the most.	Q1	Aviation Exposure (Q2)
I feel like I identify more with Navy Pilots but if I am not chosen for this I would gladly serve as a Marine Pilot.	Q3	Aviation Exposure (Q2)
I expected to have a good time bonding with my classmates and to learn more about the Marine Corps. Leatherneck matched my expectations	Q6	LNK's Impact on Service Selection (Q3)

Real World Applications

- Marine Corps Center for Lessons Leaned (MCCLL):
  - Categorizing after-action reports (AARs) in response to user search queries using Long Short-Term Memory (LSTM) networks
- Derive new sentiment dictionaries utilizing Machine Learning
  - Amazon Comprehend
  - Google Cloud Natural Language
- “KNN based Machine Learning Approach for Text and Document Mining”
  - http://web.inf.ufpr.br/menotti/ci171-2015-2-1/files/seminario-Fabricio-artigo.pdf

- ESPN: Fantasy Football and IBM Watson





Common Issues/Concerns

- Sentiment scoring using non-English words
  - Emojis: ☺, ☹
  - Slang terms: lol, haha, jk
  - Non-English language (Arabic translations)
- Negation
  - “The platoon was never late.” vs. “The platoon was always late.”
  - Some text mining software can identify negation relatively well; however, more sophisticated approaches might be necessary (i.e. N-gram sentiment scoring).

Easy to Use Text Mining Packages in R

- tidytext
  - Data structure: Tibble (fancy data frame)
  - Ideal for sentiment analysis and word groupings in a single document, can be scaled if necessary
- text2vec
  - Data Structure: Document Term Matrix (DTM), rows correspond to documents and columns correspond to a term (word)
  - Ideal for similarity comparison due to its vectorization capabilities
- tm
  - Data structure: Corpus (collection of text documents, term-document matrix, bag-of-words)
  - Ideal for large processing of text documents

Questions/Comments