Feature Extraction from Textual Datasets

Patrick Moran
Bethany Herwaldt, Jeffrey Salter
Shaina Race, Ralph Abbey
Carl Meyer
The Problem

- We have a collection of related textual documents.
  --- Ours were product reviews of a Leica DLux camera.
- We want to identify the topics being discussed.
  --- Weight, picture quality, bells-and-whistles, etc.
- We want to judge the positivity or negativity of opinions being expressed.
  --- This is future work.
Outlined Approach

- Pre-process our input.
- Create a relatively short list of “topic words.”
  --- Words likely to pertain to a specific topic.
- Generate a graph of relationships between these topic words.
  --- How related are two words to each other?
- Cluster these words together.
  --- Each cluster should be interpretable as a topic.
Pre-Processing

- Stemming
  --- “run”, “running”, “runner” all become “run”.
  --- Used libstemmer (started by Dr. Porter)
Non-Negative Matrix Factorization

\[ A \approx WH \]
NMF - Interpretation

- Each column of the approximation WH is a linear combination of the columns of W.
- The weights of these combinations are given by the columns of H.
- We can interpret this as a soft-clustering of the documents.
  --- Each column of W is a prototypical document for a given topic.
  --- Actual documents are a linear combination of topics.
NMF - Interpretation

- Some garbage topic vectors are expected.
  --- There is a great deal of “normal” English filler in a sentence.
  --- Articles, for example, will appear in all documents.
  --- This is, essentially, the noise in our dataset.
NMF – Algorithmic Concerns

- We used Patrick Hoyer's NMF with sparsity constraints.
  --- Enforced sparsity, improving the interpretability of the results.

- Empirically, the success seems pretty independent of the rank of approximation.
  --- More on this in a minute.
NMF – Results

- noise, buy, sensor, panasonic, silly, fuji
- quality, manufacture, pay, operational, lens
- format, shoot, flash, slowlag, promise, automotive, flashoth, side, equipment, inside
- image, color, clarity, small, size, alternative, mk, lightweight, sturdy, c-lux
- camera, amazing, happy, menu, master, photo, mp, close
NMF – Results

- noise, buy, sensor, panasonic, silly, fuji
- quality, manufacture, pay, operational, lens
- format, shoot, flash, slowlag, promise, automotive, flashoth, side, equipment, inside
- image, color, clarity, small, size, alternative, mk, lightweight, sturdy, c-lux
- camera, amazing, happy, menu, master, photo, mp, close

Sensor/Lens
NMF – Results

- noise, buy, sensor, panasonic, silly, fuji
- quality, manufacture, pay, operational, lens
- format, shoot, flash, slowlag, promise, automotive, flashoth, side, equipment, inside
- image, color, clarity, small, size, alternative, mk, lightweight, sturdy, c-lux
- camera, amazing, happy, menu, master, photo, mp, close

Generic Camera Words
NMF – Results

- noise, buy, sensor, panasonic, silly, fuji
- quality, manufacture, pay, operational, lens
- format, shoot, flash, slowlag, promise, automotive, flashoth, side, equipment, inside
- image, color, clarity, small, size, alternative, mk, lightweight, sturdy, c-lux
- camera, amazing, happy, menu, master, photo, mp, close

Alternative Cameras
NMF – Results

- noise, buy, sensor, panasonic, silly, fuji
- quality, manufacture, pay, operational, lens
- format, shoot, flash, slowlag, promise, automotive, flashoth, side, equipment, inside
- image, color, clarity, small, size, alternative, mk, lightweight, sturdy, c-lux
- camera, amazing, happy, menu, master, photo, mp, close

Size / Weight
NMF – Results

- noise, buy, sensor, panasonic, silly, fuji
- quality, manufacture, pay, operational, lens
- format, shoot, flash, slowlag, promise, automotive, flashoth, side, equipment, inside
- image, color, clarity, small, size, alternative, mk, lightweight, sturdy, c-lux
- camera, amazing, happy, menu, master, photo, mp, close

Image Quality
NMF – Results

- noise, buy, sensor, panasonic, silly, fuji
- quality, manufacture, pay, operational, lens
- format, shoot, flash, slowlag, promise, automotive, flashoth, side, equipment, inside
- image, color, clarity, small, size, alternative, mk, lightweight, sturdy, c-lux
- camera, amazing, happy, menu, master, photo, mp, close

Garbage / Unknown
What Happened?

- Using NMF for soft clustering assumes that related words co-occur.
  - With many, very short documents, related words are often alternatives.
  - These can even be less likely to co-occur than average, which certainly invalidates this assumption.

- Nonetheless, we do get lots of good topic words.
  - We want to filter the bad ones.
  - We want to group them.
Filtering Words

\[
\frac{f_{di}}{f_{Ei}}
\]

- Divide frequencies of each word in your dataset to their frequency in the “English language.”
  - The “English language” is some large corpus of English text.
  - We used TV and movie scripts.
- The higher this ratio, the more uncommonly-often a word is used.
  - Words with higher ratios are more likely relevant to the subject field.
Combining Metrics

- Using only word-usage ratios gives misspellings high weight, as they are “rare” in English.
- Simply using words from the NMF gives overly common words.
  --- However, the top word of each column was always good.
  --- Usually dominant by a factor of 2 – 10.
- Filtering NMF words with word-usage ratios allows us to use only words that are likely by both metrics.
Graphing the Keywords

- Now we have a list of topic words.
- We define a graph.
  --- The distance between two nodes is a measure of how similar they are.
- Similarity is based on two factors.
  --- Semantic Similarity
  --- Word Proximity
Semantic Similarity

- A measure of how close two words are in meaning or relatedness.
  --- Independent of context.
- This allows us to group words related to each other, even if they don't appear near each other.
- Necessary because we generally have low-quality data.
  --- It serves to amplify signal to help overcome noise.
Semantic Similarity - WordNet

- Heavy
- Large
- Massive
- Weight
- Big
- Size
- Little
- Largeness

Relationships:
- Attribute Of: heavy -> weight
- Similar To: massive -> big
- Synonym: massive -> big
- Hyponym: size -> little
- Antonym: big -> little
Semantic Similarity – Finesse

- After the subgraphs meet, we go one iteration further.
  --- We then take the size of the overlap as a second metric.
  --- Words could be related through obscure meanings.

\[
S_{i, j} = \frac{d_{i, j} - 1}{5} + \frac{20 - o_{i, j}}{20}
\]
Word Proximity

- The frequency with which words appear close together.
- This allows us to infer which words are related by the context in which they appear.
Word Proximity

- Cui, Mittal, and Datal concluded that there is no significant relationship between words more than 5 apart.
- For each pair of words, we count up the number of times they appear within 5 words of each other.
- We divide this by the min of the number of occurrences of the 2 words.
Clustering the Graph

- The graph distance is some linear combination of semantic similarity and word proximity.
  --- Empirically, even weighting did well.
- Then, we associate together the words with the strongest relationships.
The Graph
Clustering Into Topics

- Each cluster should be a topic.
  --- Words related by context, meaning or both.
- Any graph-clustering algorithm can be used.
  --- We projected the data into a lower dimensional space via an SVD, then partitioned with Principal Direction Gap Partitioning (PDGP), then post-processed with K-means.
  --- Unfortunately, no theory for selecting the number of clusters.
Results - Good

--- image, images, color, quality, clarity
--- lens, optics, image, sturdy
--- canon, nikon, sony, mp, packaging
--- pictures, candid, landscapes

--- options, menu, item, manual, settings, sensor, photographer, worlds, shoots
--- love, very, great, also, expensive
--- camera, cameras
Results - Bad

- use, its
- Delicate, shipping, raw, mode, ratio
- size, post, noise, flash, screen
- feature, format, shoot lightweight
- everyday
- grandchildren
- aspect
- digital, compact, complicate, swears
Drawbacks and Limitations

- As always, selecting the number of clusters is tricky.
  --- Empirically, selecting the wrong number could give very poor results.

- There are a lot of parameters.
  --- Most have reasonable default values, but some do not.

- Results are far from perfect.
  --- Definitely better than random.
Area of Improvement – NLP

- It would help to replace word proximity with some measure of word relatedness.
  --- This would require some natural language programming to implement.
- There are an awful lot of complexities.
  --- Pronouns within sentences
  --- Pronouns across sentence boundaries
  --- Type of speech detection
  --- Misspelling, bad grammar
Area of Improvement – WordNet

- Currently, we treat all word relationships equally.
  --- Synonym should probably be closer than hyponym.
  --- One would need to consult with a linguist.

- Patterns of word relationships might add or subtract weight.
  --- hyponym – hypernym
    - This goes “up” in genericicity, then back down.
Area of Improvement – Corpus

- The corpus of English text could be refined.
  --- Removal of confirmed misspellings
- The English corpus could also be expanded.
Alternative Approach – Hard Clustering

- Don't form the columns of A from documents, but from sentences.
  --- Then a “document” probably consists of only one topic.
  --- A more traditional hard clustering can be used on the sentences.
  --- We must normalize and weight the sentences to avoid long reviews automatically being given preference over short ones.
  --- This would produce many more garbage clusters, but hopefully also better topic clusters.
Conclusion

- We start by trying to identify words which characterize various topics.
- We then build a graph of these words, based on word relatedness metrics.
- Finally, we cluster this graph to arrive at a set of topics.
- This algorithm does seem to work, but has room for a lot of improvement.