NLP in business

ZHENYA ANTIĆ
PRACTICAL LINGUISTICS, INC.
APRIL 16, 2019

About

- Independent consultant and founder of Practical Linguistics, Inc.
- Previously: FactSet Research Systems
- PhD in Linguistics from UC Berkeley, BS in Computer Science from MIT
- Contact: zhenya@practicallinguistics.com
- https://www.linkedin.com/in/zhenya-antic/
  - #languagenlpforbusiness

Topics

- Marketing
  - Customer reviews and social media analysis
- Competitor analysis
  - Custom news reports
- Customer service and marketing
  - Chatbots
- Research
  - Computer generated book
- Finance
  - Alpha signals from social media

Marketing: customer reviews and social media analysis

- Social media, reviews: important marketing and feedback tool
- All kinds of businesses rely on it
- There are important insights in the data, but not readily available

> Don't let the pizza parlor storefront or steep, narrow flight of stairs put you off, this place really knows how to do homemade Italian and **the price is right**! BYOB, not too crowded on a Friday night, **great service**, and **very good food** (special attention to the homemade pastas and sauces- pappardelle and black squid linguini were best). All in all, looking forward to returning!! [Emphasis mine]

Marketing: customer reviews and social media analysis

- Main points of the review: good price, great service and very good food
- It would be great if we could collect all reviews and social media mentions for a business and extract aggregate data about each topic (price, service, food)
- Good for businesses with a large number of reviews that are hard to go through manually

Advantages of such an analysis

- Clarity and efficiency: review snippets grouped by topic
  - Analyze sentiment by topic
- Insights over time: possible to see changes in customer sentiment
- What to do with the insights
  - Emphasize in marketing materials (“we have the best customer service”)
  - Justify certain sentiment (talk about quality ingredients where the food price might seem high)
  - Correct problems

How to do it?

- Using unsupervised learning (almost)
- First, we need a domain model
  - In this case, restaurants
  - Collect lots of reviews about restaurants and use as a base
  - Split each review into sentences (and sentence parts)
  - Do lots of preprocessing
  - Try to cluster the sentence bits using K-Means
  - And…?

How to do it?

- Any ideas?
- We used word2vec to enrich the sentence bits, adding similar words
- Use TF-IDF, and then cluster using K-Means

Clustering code

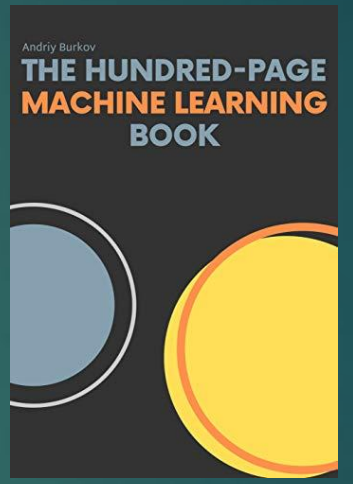
```
add_t = model.most_similar([word], topn=15)
add_w = [m[0] for m in add_t if word not in stopwords]
vec = TfidfVectorizer(max_df=0.90, max_features=200000,
    min_df=0.05, stop_words=stopwords,
    use_idf=True, tokenizer=tokenize_and_stem,
    ngram_range=(1,3))
tfidf_vectorizer = vec.fit(enriched_strings)
tfidf_matrix =  tfidf_vectorizer.transform(enriched_strings)
km = KMeans(n_clusters=num_clusters, init='k-means++',
    max_iter=300, n_init=10, random_state=0, verbose=0)
km.fit(tfidf_matrix)
```

Using the model

- Once the model is built, some calibration will be necessary
- Experimenting with the number of clusters
- Naming the clusters
- Stopwords addition/removal: domain specific
  - Company names
  - Words like “good”, “great”, etc.
- Then, using new sentence bit, return the cluster it belongs to

Experimenting with number of clusters

- To determine n automatically:
- Split data into training and test sets
- Do the training clustering and the test clustering
- For each two points in the test set that are in the same cluster, check if they would be in the same cluster in the training clustering
- If n is an optimal number, the training and test clustering should be consistent
- Prediction strength = proportion of pairs that are in the same cluster in both training and test clustering
- Pick n where prediction strength > 0.8

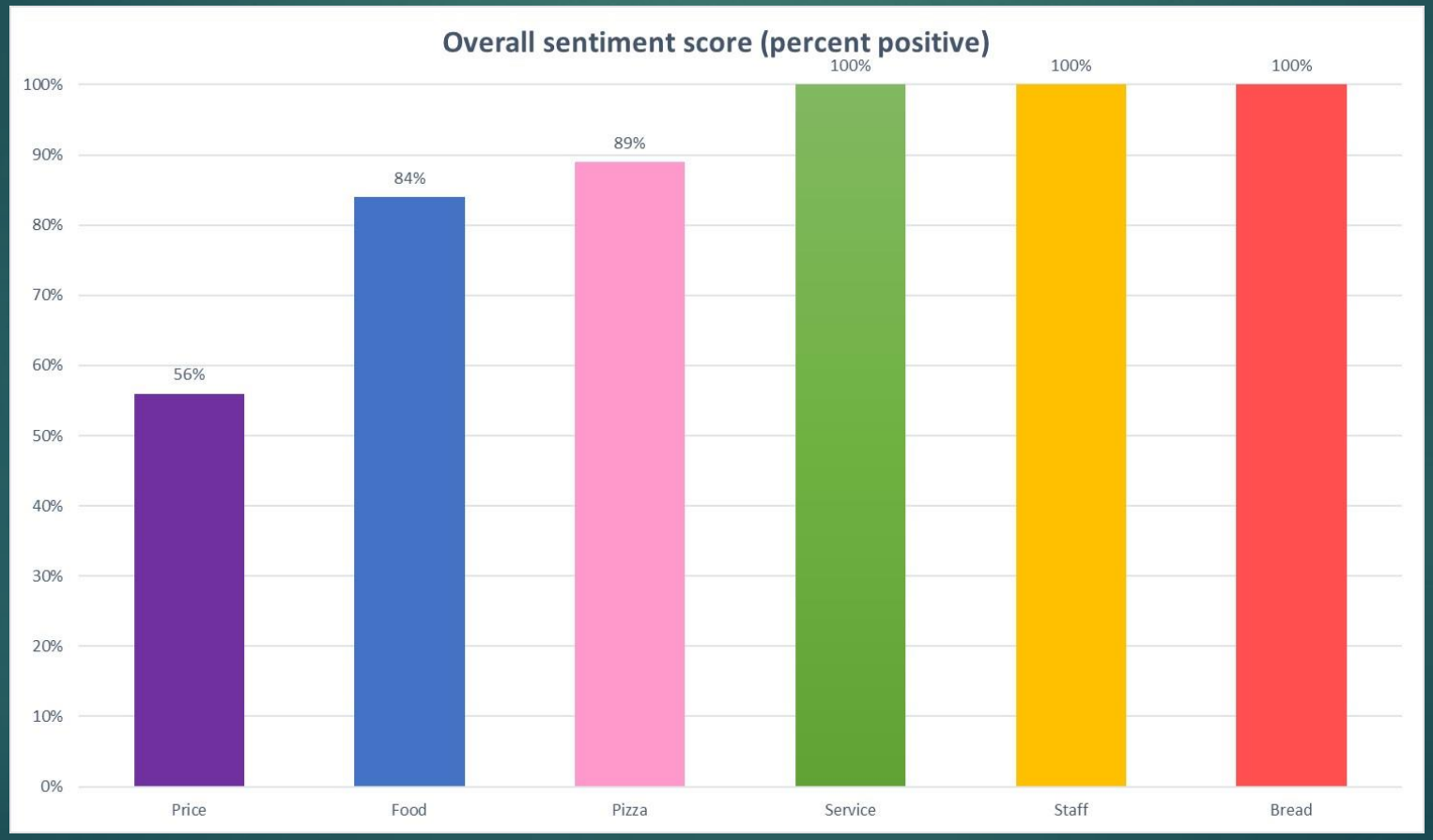


Restaurant model

- Don't let the pizza parlor storefront or steep, narrow flight of stairs put you off, this place really knows how to do homemade Italian and **the price is right**! BYOB, not too crowded on a Friday night, **great service**, and **very good food** (special attention to the homemade pastas and sauces- pappardelle and black squid linguini were best). All in all, looking forward to returning!! [Emphasis mine]

- Topics: service, staff, food, price, pizza and bread (and others, discarded)

Restaurant model



Restaurant model

- **Service. Overall sentiment score: 100% positive**
- *positive* “service was good”
- *positive* “service has always been good”
- *positive* “the service was great”
- *positive* “service was pretty decent!”
- *positive* “so service is really good”

Restaurant model

- **Staff. Overall sentiment score: 100% positive**
- *positive* “and friendly staff”
- *positive* “the staff in the kitchen are very nice people”
- *positive* “the staff is super friendly”

Restaurant model

- **Food. Overall sentiment score: 84% positive**
- *positive* “not only is the food amazing and always up to par”
- *negative* “2/5 because bad food”
- *positive* “the food here is amazing!”
- *negative* “the food was n’t great either.”
- *positive* “food is fresh and steaming hot.”

Restaurant model

- **Price. Overall sentiment score: 56% positive**
- *positive* "if you want quality for a decent price, Pomodoro is a way to go"
- *negative* "a price that is a bit high for what I got"
- *positive* "fresh homemade pizza for reasonable prices"
- *positive* "did my son's birthday upstairs for a very reasonable price"
- *negative* "Although it's a bit pricey"

Restaurant model

- **Pizza. Overall sentiment score: 89% positive**
- *positive* "the pizza was great"
- *positive* "we got the 16 "abruzzo pizza and it was wonderful!"
- *negative* "we stopped bothering getting pizza there because it is a little more expensive than donnagio's and not much better"
- *positive* "their pizza is probably one of the best in the area if you like the more thin crust pizza"

Restaurant model

- **Bread. Overall sentiment score: 100% positive**
- *positive* “their freshly baked bread was divine!”
- *positive* “the only good part was the bread”
- *positive* “the bread they bring out for free is amazing!”

Customer reviews

Customer reviews: fake reviews

- Fake reviews
- People are bad at identifying fake reviews
  - Even people who write them
- Labeled datasets are hard to find
  - Constructed datasets, but…
- Analyzing the text for unusual features: extremely positive, extremely negative, keyword stuffing (mentioning the brand too many times), etc.
- Analyzing user behavior: ongoing marketing campaigns – writing reviews in bursts, very similar reviews, etc.
- Keep up with the latest research

Custom news reports

- News can be a valuable resource for companies
  - Reports about events in the industry
  - News about competitors
  - Monitoring news about themselves
- Small coffee house chain
  - Global news: coffee, supplies prices that might affect them
  - Local news: competitors, construction in the area where they are

Custom news reports

- Stream of incoming news
  - Some sources are available for free (a deprecated Google API)
  - Google RSS feed
  - A paid news stream
  - Scraping news from websites (beware of policies)
- Preprocessing
  - Getting text out of HTML

Custom news reports

- Classification of news by topic
  - Custom labels that will depend on the nature of business and interest (coffee prices, local construction, local events, competitors)
  - Labeled data
- Event extraction
  - Coffee price going up, new restaurant opening/closing
- Company extraction
  - Competitors
  - Primary and secondary mentions

Customer service and marketing: chatbots

- Popularity of messaging on the rise
- Still relatively little spam: not much marketing through the channel
- People like talking to bots: the ELIZA effect
- Customer service
  - Hours, address, contact information
  - Frequently asked questions
  - Collect user contact information
- Marketing
  - Audience segmentation using the data available is much easier
  - Possible to ask user questions to find out more about them (with care)
  - Provide promotional material (with care)

Chatbots

- Need to have narrow scope
  - The broader the scope, the harder it is to get the user intent
- Set user expectations upfront: let them know they are talking to a bot
  - ELIZA effect still exists
- Let the user know what the bot can and cannot do

Chatbots

- Use a platform to build a bot (there are many)
- Do it yourself
  - Can use help of platforms, such as DialogFlow (Google) and wit.ai (Facebook)
- Collect data (user questions)
  - Small bot that collects questions
- Use rule-based, ML, or neural networks to answer user questions
  - Depending on the amount of data

Research

- Summarization
  - Computers are pretty bad at it
  - Extractive summarization
  - Abstractive summarization
- Search for similar documents
- Aggregating documents by topics
- Uses: legal, healthcare

Research: computer generated book

- Lithium-Ion Batteries, A Machine-Generated Summary of Current Research
- A large set of research articles chosen by keywords
- Divide the set into chapters, chapters into sections
- Each section:
  - Introduction
  - Summaries of chosen articles
  - Conclusion

Techniques

- Chapters: K-means on the TF-IDF matrix
- Further divide chapters into sections using K-means
- With manual refinement
- Alternative: bibliography overlap, but biased against publications with large number of references

Techniques

- Section names: the most characteristic word while clustering
- E.g., “Anode Materials, SEI, Carbon, Graphite, Conductivity, Graphene, Reversible, Formation”
- They tried a neural network method, but found it hard to get consistent quality

Techniques

- Introduction, conclusion: concatenation of summaries of all document introductions
- Document summaries: extended abstracts
  - Augment the paper abstract with sentences from the body by using n-gram overlap similarity
  - Linguistically reformulate sentences
    - Rule-based simplification (e.g., remove sentence initial adverbials)
    - Sentence compression (e.g., remove local/temporal cues)
    - Sentence restructuring (e.g., active -> passive)
    - Sentence reformulation (substitution of synonyms, something like word2vec)
    - Anaphora resolution

Based on transition metal oxides (TMOs) including TiO_2 [67], ZnO [68], CuO [69], Fe_3O_4 [70], NiO [71], CoO_x [72–75] as anode materials for Li-ion batteries, and MnO [76], has made considerable progress among the wide range of efforts [4]. Co_3O_4 materials with multiple structures have been efficiently prepared, including lamellar [77, 78], nanorods [79], hollow spheres [80], nanoparticles [81, 82], and cubes [4, 83]. High lithium storage Co_3O_4 electrodes could be obtained by the indicators of designing hollow structures [4]. There is still a challenge to enhance the electric conductivity and agglomeration issue of Co_3O_4, which are the contextual factors impeding the development of Co_3O_4 electrodes for use in Li-ion batteries [4]. Carbonaceous materials have functioned as the most optimum conductive materials to enhance the electric conductivity of Li-ion batteries' electrodes [4]. Two-dimensional (2D) graphene (GR) with an excellent electric conductivity, systemic flexibility [84], and rich surface area, is another influential carbon material [4]. A hybrid of these two types of materials which formed a new 3-D (3D) layered structure is the most efficient technique in order to harness the advantages of the 1D CNTs and 2D GR [4]. The 3D graphene/carbon nanotubes (GR/CNTs) network can not just maintain the excellent properties of CNTs and GR though enhance the inferior electric conductivity between graphene sheets [4, 85]. Co_3O_4 hollow microsphere/graphene/carbon nanotube (Co_3O_4/GR/CNT) flexible film is prepared through a two-stage technique; this technique comprises a subsequent thermal decrease process and a straightforward filtration route [4]. That the film electrode showed better lithium storage capacities in rate and cycling performances than hollow Co_3O_4 materials is revealed by the results [4].

Numerous researches on CuO/graphene composites utilized as Li-ion batteries anode have been indicated; for instance, Rai and others [86] have synthesized CuO/rGO nanocomposite through a spex-milling technique [5]. The first discharge capacity of 1043.3 mAh g^{-1} had been delivered by the CuO/rGO composite, and the charge capacity can be maintained at 516.4 mAh g^{-1} after 45 cycles at 0.1 mA cm^{-2} [5]. Enhanced anodic performance, which is compared to the pure CuO nanoparticles, had been shown by this CuO/rGO composite [5]. A novel kind of CuO nanosheets/rGO composite paper, which revealed better cyclic retention than that of the pure CuO nanosheets had been indicated by Liu and others [5, 87]. Improved electrochemical performance than pure CuO had been demonstrated by the composites [5]. Porous CuO nanorods/rGO had been synthesizeded by Zhang and others [88] composite through hydrothermal reaction [5]. Improved electrochemical properties than the pristine CuO nanorods were shown by the composite electrode [5]. A facile refluxing approach had been utilized to synthesize ultra-short rice-like CuO-NRs/rGO composite [5]. Cu^{2+} ions absorbed into $Cu(OH)_2$ and then rapidly dehydrated into CuO-NRs under high temperature, with homogeneous distribution on the rGO nanosheets after the addition of NaOH [5]. The as-prepared CuO-NRs/rGO composite anode indicates enhanced electrochemical performance in Li-ion batteries due to the synergetic effect between the high electrical conductivity of rGO nanosheets and the well-dispersed CuO-NRs [5]. The rGO

Finance: alpha signal from social media

- Jamie Wise from Periscope at the AI and Data Science in Trading
- Collect tweets about companies
- Label them positive, negative or neutral
  - Neutral is important
- Use them to predict company performance

Other uses

- Recruiting
  - Candidate recommender systems
  - Ideal profiles for candidates
  - Information extraction from job descriptions and resumes
  - LinkedIn displays information such as “You are in top 10% candidates for this job”

Thank you