Final Capstone Product

Build a Model that can be considered a “Data Science Product”.

It should solve a real problem or create something of value.

Deliverables:

* A one-page proposal to be approved by the data science team here, before you begin
* A 20 Minute Presentation on your product that argues not only for its relevance, but why this implementation is best, and the technical tools you will use to get there. This presentation should include a demo!
* A Jupiter notebook that builds the model and shows your machine learning tuning process
* Defend your model in a 60-minute mock interview (the interviewer will attend your presentation and read through your notebook beforehand)
* A final 60-minute mock interview on data science generally

Data Science Requirements:

* Utilize at least one specialization in your pipeline
  + Advanced Natural Language Processing and Neural Networks
* Include supervised and unsupervised techniques

Guidelines:

* Your data set will be hugely important here. Though we've relied on pre-aggregated data before, for this assignment gathering your own data through scraping or other techniques may allow you to make a more impactful or novel product. It will likely need to see a continuous flow of data in order to be relevant as a data solution.
* Emphasize why your implementation is the best implementation, not just an implementation
* Remember, a good product is more than just metrics. It solves a problem.
* You do not have to make a front end or a user interface for this product

Ayan Karim Proposal

01/14/2019

**Opinion Mining using Aspect-Based Sentiment Analysis for Analyzing Competition and Product Development**

**Problem**

Tech companies want to assess their products’ performance and their consumer’s opinions to their products. Most companies have their own forum for complaints and methods of collecting feedback, but most people don’t place general reviews and opinions directly on those forums unless they feel strongly about something specific that’s affecting them, so there’s a selection bias.

Furthermore, they want to understand how their product compares with that of their competitors, not just in sales but in performance.

To get the whole picture of how consumers feel about their product as well as that of their competition, the company needs information from public media like news and social media.

The **problem** is that there’s an opportunity cost in the time spent on researching reviews online and deciding which reviews are meaningful to a company for product development, then repeating that research on competitors for the sake of comparison.

**Value**

My product will do the following:

1. Collect consumer opinions on a company’s product and that of their competitor from popular media sources online.
2. A semantic analysis of the reviews and categorize their semantics by positive, negative or neutral.
3. It will cluster the text into topics (memory, speed, user-friendliness, reliability, etc.) that are related to positive and negative sentiments and sort the aspects (terms that refer to one of the topics) under each sentiment, based on popularity.
4. The results will return two queries, one on your companies’ opinions and one on your competitors.
5. Finally, my product will produce visualizations that compare the two companies’ opinion data based on their positive and negative sentiments and how they differ in each topic.

The **value** that my product gives is return an easy to read and visualize query of popular opinions on a company’s product as well as which aspects the opinion refers to. And it compares these opinions to that of their main competitor. A company can then use this information for product development.

**Data Source**

The data will be textual information scraped from TechRadar, Wired and Twitter.

**Techniques that will be used in this project**

* Natural Language Processing
  + Spacy for text processing
  + Sense2vec and TF-IDF Vectorizers for feature engineering.
  + LSA and its variants for topic modelling.
  + TextRank for popularity and significance
* Linear Support Vector Classifiers
  + Classification approach to topic modelling.
* Logistic Regression
  + May use a pre-trained model to use classify sentiments
* Neural Network
  + Neural Coeff v2.0 for pretrained neural networks that parse and replaces pronouns

**Possible Challenges**

* My biggest challenge will be scraping the data and cleaning it into a form I can process and explore.
* Figuring out how to analyze the semantics of textual data that don’t have labels. I’ll probably use a pretrained model.
* Producing the resulting tables for each company so that they’re easy to compare side by side and produce the accompanying visualizations.

1. Scrape Data from Twitter
2. Scrape Data from TechRadar
3. Pre-Process Data