Sentiment analysis on Amazon reviews

Helen Pung, Samuel Johannes Pitko

Business understanding

- Identifying your business goals
    - **Background** - This project's story begins on a monday morning, about 3 hours before the deadline for the project introduction slides. Both our team members were desperately searching for a topic to soon present. Given the time restriction they decided to look around on Kaggle in the dataset section for something that hadn't already been selected by 10 other teams. It was then that they landed and stuck with an intriguing dataset called Amazon Cell Phones Reviews because it would allow them to explore something that had not been previously covered in practice sessions - Natural Language Processing and sentiment analysis.
    - **Business goals** - Our primary business goal with this project is to provide a possibility for extracting a reviewers polarity expressed in their review. Although our chosen target data already includes a metric for this - star rating, it could be possible to use the same model to predict the polarity of texts with similar contexts, for example reviews on other sites without ratings, in order to gain valuable insight into different products or brands.
    - **Business success criteria** - This project will be considered a success should we be able to develop a model capable of reliably classifying a review as either negative or positive. Additionally, if we're able to find and visualize at least two different interesting visualizations of the given data, that will also be considered a success.

- Assessing your situation
    - **Inventory of resources** - The list of people available for this project is limited to the two members of this project's group. For data however, there is plenty of datasets about reviews with included ratings available, including two smaller instances on Kaggle. Should the need arise for more data, there is multiple years worth of review data available straight from Amazon.
    - **Requirements, assumptions, and constraints** - Requirements for this project to succeed include sufficient available data, tools and hardware. Should there be a need for additional, more powerful hardware, this could be acquired from Kaggle, Google or similar services.
    - **Risks and contingencies** - This project does not pose any major risks directly related to the project itself. Obviously general best practises will be applied as much as possible in order to prevent any common issues, e.g using version control to avert data loss.
    - **Terminology** - Our project's terminology will be limited to standard statistics and visualization terms.
    - **Costs and benefits** - This project's only cost is only the precious time of two busy students. As such it should be evident that the benefits gained, including passing this course, will exceed the project's costs.

- Defining your data-mining goals
    - **Data-mining goals** - Primary goal is to develop a model to predict the polarity (negative/positive) of a review using either statistical modelling or machine learning. This includes exploring, understanding and preprocessing/cleaning the supplied data. During this process, we will also attempt to find at least two interesting insights that could be well visualized.
    - **Data-mining success criteria** - In order to consider this project successful, the developed model's accuracy should exceed 0.85, hopefully reaching 0.9.

Data understanding

- ## Gathering data
    - **Outline data requirements** - As our project's topic was chosen based on the dataset, there should not be any additional types of data necessary for this project to succeed. The two chosen datasets are both provided in comma delimited files and include data mostly from the last 5 years.
    - **Verify data availability** - Considering Amazon itself provides this same data, there should not be any issues with data being unavailable or unaccessible.
    - **Define selection criteria** - We will be using the two datasets found on kaggle. Most relevant columns for this project are review rating, title and body, although some other columns, for example the amount of "helpful" votes could also provide valuable information. This will need to be further explored.
- ## Describing data

  Both of the selected datasets contain data about Amazon product listings' reviews. Columns include information about the product itself - name, price and the like, the reviewers name, verification status and details about the review - rating, title, body, amount of "helpful" votes received.

- ## Exploring and verifying data

  Total row count of selected datasets is about 480,000. Ratings are provided in a range of 1-5, representing the number of stars of the review on Amazon. Both of the datasets seem to be of high quality with little to no major problems. Data preparation will mostly consist of converting the text itself into a machine readable format, be that word tokenization, vectors or something similar.

Planning

Task	Tools	Time	Notes
Embracing the Data-Mining Process	Text editor	Samuel: 3h Helen: 1h	Data and Business understanding, planning
Research available tools and frameworks	Google, Kaggle, Github or similar	Samuel: 2h Helen: 4h	
Develop a predictive model	Pandas, numpy	Samuel: 15h Helen: 10h	Keras in case of using ML
Visualize insights found from data	Matplotlib, seaborn, pandas	Samuel: 5h Helen: 10h	Wordcloud
Make an appealing and informative poster	Photoshop, Illustrator or similar	Samuel: 4h Helen: 6h	