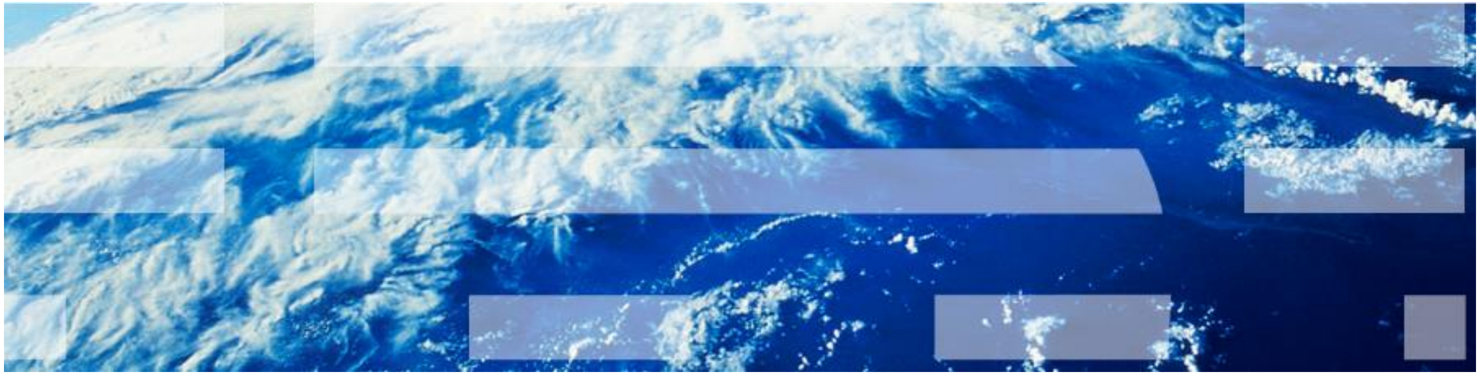
E6893 Big Data Analytics

Arbitrary Aspect Identification, Extraction, and Ranking

Project ID: 201912-36
Team Members (with UNI): Cedric Jouan (cj2567) and Austin Bell (alb2307)

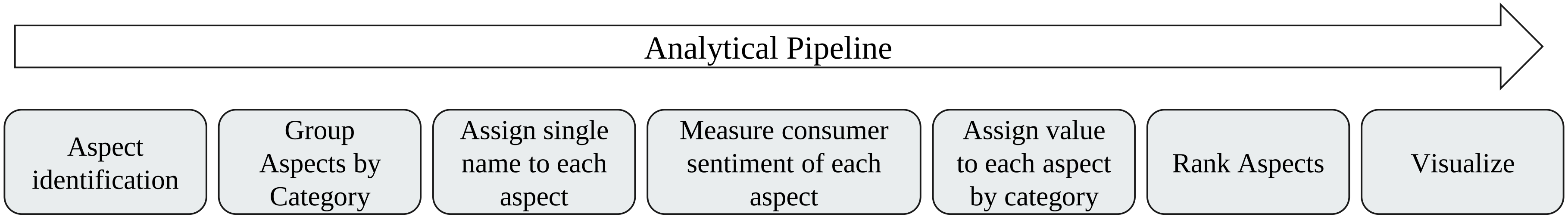


Intro and Goal

Goal: Identify, extract, and rank the importance of product aspects for each category on Amazon

- Aspect refers to a component or attribute of a specific product (e.g., an aspect of a phone may be its battery life)

Leveraging the unstructured text in Amazon Product reviews, we can identify aspects for different categories and measure consumer sentiment towards each aspect.



Example: Aspect Identification

Review of Iphone 11

Nina

Attractive looking phone

October 26, 2019

Size: 64GB | Color: Silver

I love this iPhone so far. I first has the non pro and the screen was killing my eyes. Everything just seemed blurry so I returned it and got this pro and the screen it's clearer than ever. The design looks real good I like the 3 camera design and the bigger screen due to not having a home button. I just can't stop looking at it. I just put a matte screen protector on it and a otter box case so i'm good to go. I didn't use the fast charger yet or the ipods so I don't know how they work yet. I didn't notice a 2.5mm port on it. So I guess it doesn't have one Maybe the iPods are supposed to be attached by the charging port. I'm satisfied with my iPhone.

Potential aspects:

- Screen
- Design
- Camera
- Charger
- 2.5mm port
- Ipods

Proposed Value

Our analysis looks to better understand reasons for consumer purchases

- Which aspects contribute most to positive opinions towards specific products?
- Which aspects correlate most with likelihood of purchase within a particular product category?

Through leveraging open text product reviews, we seek to:

- Reduce potential bias of respondents found in survey analysis
- Reduce cost and time compared to controlled experiments

While simultaneously developing a framework that is highly scalable and replicable

Dataset

Amazon Product Reviews

- Product reviews from May 1996 - July 2014
- Includes: review text, rating, and product ID
- 142.8 million observations
- 24 broad product categories with many more subcategories

Metadata

- Includes: product ID, category, and price
- 9.4 million products
- Links to review data via product ID

R. He, J. McAuley. Modeling the visual evolution of fashion trends with one-class collaborative filtering. WWW, 2016
J. McAuley, C. Targett, J. Shi, A. van den Hengel. Image-based recommendations on styles and substitutes. SIGIR, 2015

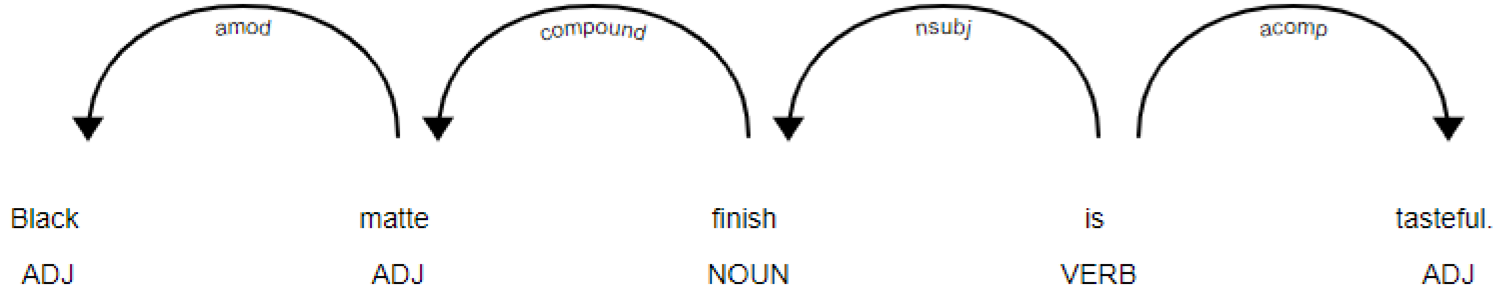
Methods

Four discrete tasks:

1. Extract aspects
2. Group and assign names to aspects
3. Compute sentiment of each aspect
4. Assign value and rank aspects

Extract Aspects

Dependency Parsing



- Product aspects are discussed in linguistically similar ways
- Defining linguistic rules and traversing dependency trees allows us to extract aspects independent of text content

Liu B., "Sentiment Analysis and Subjectivity"., **Handbook of Natural Language Processing,** 2010

Extract Aspects - Results

Result of the Aspect Extraction

Raw Review:

The battery charge is very short. Customer service is atrocious! I was on hold for 30 minutes and then disconnected. My e-mail was not returned. I had to call a local B&N for help. A 14 day return policy? Most stores give you 30 days or more. This product I believe came to the market too soon. It should've been tested more.

Extracted Text:

['The battery charge is short', 'Customer service is atrocious']

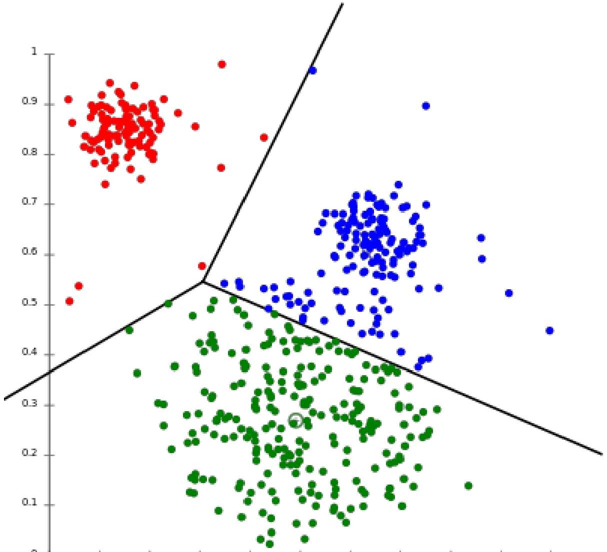
Unique Aspects:

['Customer service', 'short', 'The battery charge', 'atrocious']

Future consideration: split noun chunks and descriptors into different aspect groups

Group and Assign Names to Aspects

Clustering



- Convert extracted aspects to word vectors
- Leverage clustering algorithms to group semantically similar aspects
- Extract word(s) most representative of each group
- Assign extracted word(s) as common name for each aspect

Group and Assign Names to Aspects - Results

Initial clustering algorithm implemented:

- Assign names to aspects by main category (e.g., “Electronics”, “Toys & Games”, or “Sports”)
- Glove 100 Dimension Embeddings
- K Means random initialization, euclidean distance
- TF-IDF weighting to select most representative words

Extracted Aspect Result	Clustered
fantastic	incredible|great|fantastic
this product	use|quality|product
accurate	timely|sufficient|reasonable
children	small|similar|notarization
this software package	software|process|printing
this program	way|program|part
the price	use|quality|product
the rules	rules
a fun time	time|great|good
easy to follow	learn|easy
the rules	rules
rare	excellent

Future consideration: identify and exclude intra-cluster outliers

Compute Sentiment of each Aspect

Data Structure :

reviews	normalized_aspects	scores
It had all the songs I wanted but I had ordered the large print version and received the regular version. This was the only thing I did not like.	Size	4
I love this book. I love hymns and love to sing and run my fingers over piano/organ.This book is helpful.	content	5
We use this type of hymnal at church. I was looking for the same one; however, this wasn't it. It is a good hymnal, but there isn't enough information to find the version I need.	content	4
Heavenly Highway HymnsI ordered this hymnal because I learned to read shaped note music when I was a teenager. I play piano but do not sing. I am 85 years old. This hymnal has mostof the songs I have learned over the years. It was exactly what I wanted and needed.It was in good condition and the price was right. I purchased this book from Amazon.	condition|price	4
I bought this for my husband who plays the piano. He is having a wonderful time playing these old hymns. The music is at times hard to read because we think the book was published for singing from more than playing from. Great purchase though!	hard to read	5
This is a large size hymn book which is great to be able to see the songs, notes, etc. Quality was great, item was new!	size|condition|condition	5

Sentiment analysis :

- Split the dataset,
- Create one training set and one test set for each aspect,
- Train a Deep learning sentiment analyzer for each aspect,
- Evaluate sentiment analyzers.

$$f_{\text{aspect k}}(\text{review r}) = \text{opinion}_{rk}$$

Compute Sentiment of each Aspect - Results

Model selection and Training

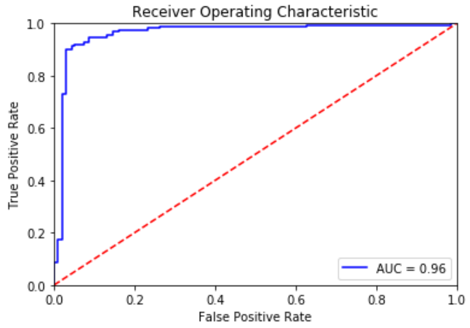
CNN_BiLSTM
Tensorflow

⊕
- Takes advantage to the large number of sample.
- Good performance.

⊖
- Prediction is slow.
- Difficult to integrate in the pipeline.

Training Data :
Texts = 500 000 unused reviews from the original dataset.
Labels = 0 if the review's score is 1 or 2, and 1 if the review's score is 4 or 5.

Glove 100 Embedding
Convolution_1D(128)
Bidirectional_LSTM(100)
Dropout(0,2)
Dense(1)



Integration in the pipeline

- **Split the reviews to isolate each aspect.**

The battery charge is very short. Customer service is atrocious! I was on hold for 30 minutes and then disconnected. My e-mail was not returned. I had to call a local B&N for help. A 14 day return policy? Most stores give you 30 days or more. This product I believe came to the market too soon. It should've been tested more.

- **Prediction.**

For each simple sentence that discuss an aspect we predict the sentiment.
Then we aggregate the sentiments in one "opinion vector" per review. E.g o_r = [0, 0, +0.9, 0, 0, -0.4]

Assign Value and Rank Aspects

- The objective is to compute the importance weights of each aspect from the opinion vectors and the overall score of the review.
- The weights should represent the influence of the aspect on the overall score.

Inputs

X= (Opinion vectors o_r) ; **y**= (scores O_r)

- The features are the sentiment score for each aspect.
- The target is the score review.

Model

We want the weight vector w such that

$$p(O_r) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(O_r - w \, . o_r^T)^2}{2^*\sigma^2}}$$

Therefore we fit a linear regression between X and y

Outputs

<u>Coefficients</u> :
The coefficients are the weights that represent the importance of each aspects.

<u>Statistics</u> :
We are also interested by the standard deviation of the coefficients.

- This model is applied at a sub category level. Therefore we output one set of weights per sub category.
- Remark : We are still developing a more robust model for the ranking.

Overview and Output description

Input dataset

Level : Full Data

```
root
 |-- index: string (nullable = true)
 |-- review: string (nullable = true)
 |-- main_category: integer (nullable = true)
 |-- category_level_1: integer (nullable = true)
 |-- scores: integer (nullable = true)
```

- 1- 142.8 million observations
- 24 broad product categories with many more subcategories

Aspects dataset

Level : Global Category

```
root
 |-- index: string (nullable = true)
 |-- review: string (nullable = true)
 |-- category_level_1: string (nullable = true)
 |-- scores: integer (nullable = true)
 |-- key_aspects: array (nullable = true)
 |    |-- element: string (containsNull = true)
```

- Create set of key Aspects for every global category
- Identify aspects discussed in each reviews

Sentiments dataset

Level : Global Category

```
root
 |-- index: string (nullable = true)
 |-- review: string (nullable = true)
 |-- category_level_1: string (nullable = true)
 |-- scores: integer (nullable = true)
 |-- key_aspects: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- sentiment_scores: array (nullable = true)
 |    |-- element: double (containsNull = true)
```

- Compute sentiments for each aspects.
- Create Opinion Vectors

Ranks dataset

Level : Category level 1

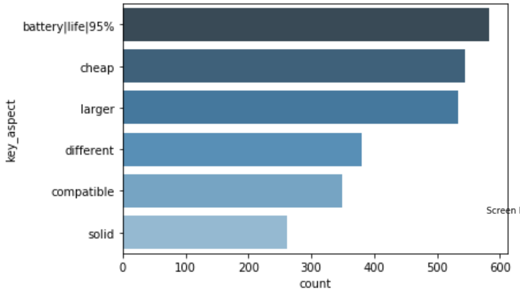
- **W** → Coefficients of the regression
- **S** → standard deviation of the coefficients
- **F** → Aspects Frequencies

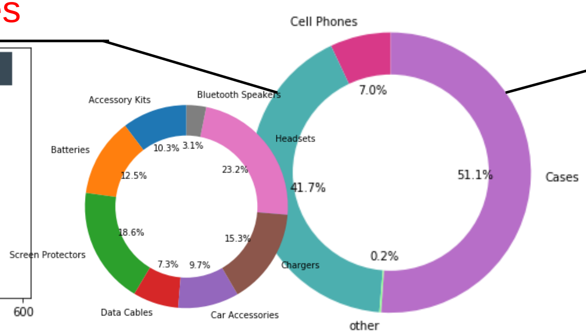
Outputs :

1. One list of aspect per main Category
2. The frequencies of these aspects in each sub categories (level 1 category)
3. The importance weights of these aspects in each sub categories
4. The standard deviation of these weights in each sub categories

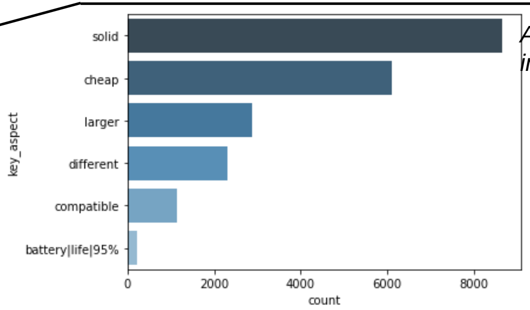
In-depth Analysis of “Cell Phones & Accessories”

Sub Category : Batteries



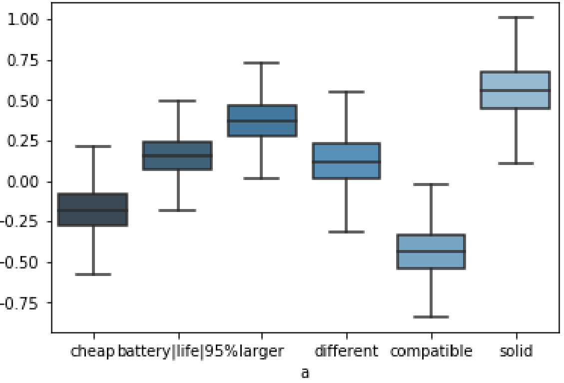


Sub Category : Cases



Aspects frequencies in the Case category

Coefficients of the logistic regression



Insights :

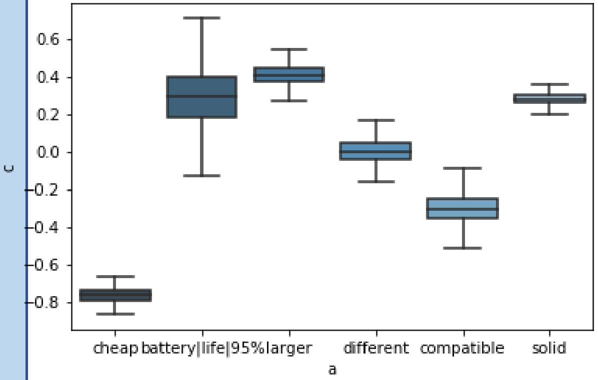
Compatibility: Most negative aspect of Cell Phones

Batterie life time: well discussed in Cell phones but seems not very important. Might be surprising.

Much less discussed in Cases category but is still important !

Price : Much more sensitive for Cases.

Coefficients of the logistic regression



Conclusion

Through chaining modern day NLP solutions, we developed a framework that provides substantial insight into consumer’s purchasing reasons

Our framework was developed such that it is independent of the data source and can be applied to any source of open text reviews in a completely unsupervised manner

Intended future work is to understand causal reasons for purchases rather than purely associative