# <noun, adjective> pair ranker

Product Review from Yelp: XXXX

Manually identified most meaningful 5 <noun, adjective> pairs:

1. Xxx
2. Xxx
3. Xxx
4. Xxx
5. Xxx

Manually added review text and rating in the csv file.

Are the top 5 ranked pairs the same as your manually identified pairs?

Most likely not.

Spacy made “best” -> “good” (lemma)

CODING

We saved the pair’s base/stem/lemma form, used their base/stem/lemma form, to calculate the frequency of the pair. This is done such that (noodles, spicy) and (noodle, spicy) will be considered the same pair.

In addition, it is observed that spacy recognise “best” -> “good” (lemma)

Therefore, we based rankings according to frequency of pairs, instead of “power” of the word,

eg Best > Good > average

## Deciding how to identify <noun, adjective>

To find the pair, we used spacy to find words with pos tag ADJ and using the dependency parsing to go “up” the tree of the ADJ, to find the noun, which pos tag is either NOUN (noun) or PROPN (proper noun)

The adjective will be paired with the nearest noun and stop. It will not be paired with other nouns that are further up the parsing tree. This is to obtain higher accuracy as pairing an adjective with multiple nouns, has a higher chance to find incorrect pairs.

<https://spacy.io/usage/linguistic-features#dependency-parse>

Example: They also come in **different** flavors **such** as chicken, carbonara, and **more**.

With Break: accuracy = =

No Break: accuracy = =

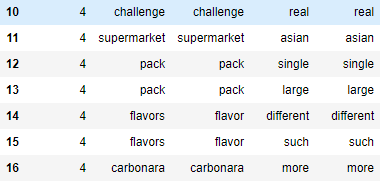
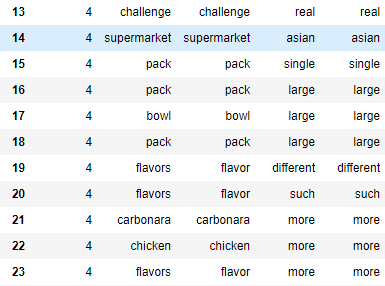
In this example, no break has worse accuracy

More = ADJ

EXAMPLE

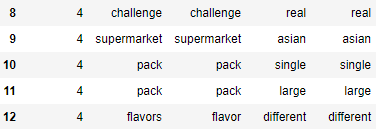
If you're looking for a **real** food challenge these noodles will give you a run for your money. They can be found at any **asian** supermarket and are very inexpensive. You can buy them as **single** pack, noodle bowl, or a **large** pack. They also come in **different** flavors **such** as chicken, carbonara, and **more**. Extremely spicy but extremely delicious.

**Has break** (only one pair) | **no break** (potential multiple pairs)

Find the syntactic dependency label = “amod”= adjectival modifier

<https://universaldependencies.org/u/dep/amod.html>



OPINION:

LEFT: 7, 2 wrongs

(flavor,such) and (carbonara, more),

5/7 = 0.714 accuracy

RIGHT: 11, 5 wrongs,

(flavor,such) and (carbonara, more), (chicken, more), (bowl, large), (pack, large)

6/11 = 0.545 accuracy

|  |  |  |
| --- | --- | --- |
| **method 1** Has break | **method 2** No break | **method 3** amod |
| lesser pairs in general | more pairs in general | least number of pairs |
| found pairs are more accurate | pairs have a higher chance to be wrong  most inaccurate | most accurate |
|  | find pair that are “further” away from each other | generally, an adjective is paired with a noun |
| pairs found = 87 | pairs found = 114 | pairs found = 78 |
| noun\_adj\_pairs\_has\_break.csv | noun\_adj\_pairs\_no\_break.csv | noun\_adj\_pairs.csv |

Dataframe **df\_pairs\_rank** 🡪 number of rows == number of unique pairs

The details of the pairs are saved in the file: noun\_adj\_pairs.csv

~~Program token words, so words are single word, noun of >1 word e.g. “chicken rice” will be split into “chicken” and “rice”~~

~~The adjective “good” will be paired with both noun “chicken” and “rice”.~~

~~However, is “good” referring to the entire dish “chicken rice”, or is “good” referring to “chicken” or “rice” individually? We are too noob to know. D:~~

## Pair frequency / program ranking

Due to small dataset ~~and reviews are generally short sentences~~, the top pair is (food, spicy) with 6 frequency, followed by ……….

As such, instead of strictly showing the 5 top pairs, we consider pairs with frequency from 6 to 2, and compared these to our manually chosen top 5 pairs.

# classify the sentences into positive, neutral, and negative classes

Textblob, intro this lib

<https://textblob.readthedocs.io/en/dev/advanced_usage.html#sentiment-analyzers>

for each review, the text is an input to the Textblob sentiment function and obtain float polarity value.

Max = most positive = 1.0

Min = most negative = -1.0

range of neutral? Neutral = -0.01 to 0.01

colours: positive = green, negative = red, neutral = blue

INSERT GRAPH

Out of the 30 review:

negative 6

neutral 2

positive 22

out of 30 reviews, ratings:

1 3

2 2

3 8

4 6

5 11

## OBSERVATION OF SENTIMENT PLOT

A pattern can be observed, a review with higher rating (e.g. 5) tend to have a higher positive value (near 0.5)

## Explain anomaly

<https://textblob.readthedocs.io/en/dev/api_reference.html#module-textblob.en.sentiments>

Example: didn't really taste good

We humans can tell that this sentence is generally negative, however the function recognise this as positive sentence, with a value of 0.449, which is relatively a high positive value.

Issue/reason is that the library/function did not take into account of “didn’t”, it only consider “really taste good”, thus resulting in a high positive value.

What are the challenges your group has encountered in this task?

30 datasets are very restrictive.

Review text is short, so find less pairs per review, e.g. average 2 pairs per review. This affects the total frequency of pairs.

average pairs / review = total number of (NOT distinct) pairs / 30 = 78 / 30 = 2.6 pairs per review

in addition,

freq, no of pairs

5 1

3 2

2 9

1 49

Sentiment analysis, we cannot train a model due to small dataset, we had to use a model based on another dataset, movie reviews.

Thus, the results may not be desirable or accurate. ~~Example: movie will not talk about taste of food.~~