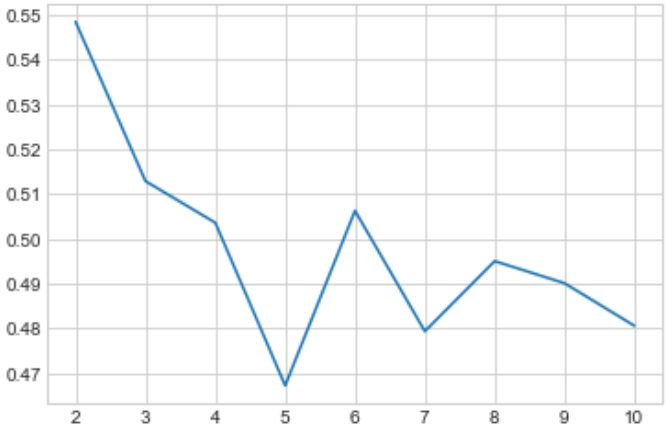
TopicModelling results summary

Common for all subreddits: Number of topics = 10, 5 top words per topic

Espresso
Comments ~ 10,000

Lda coherence = 0.44192878670786656

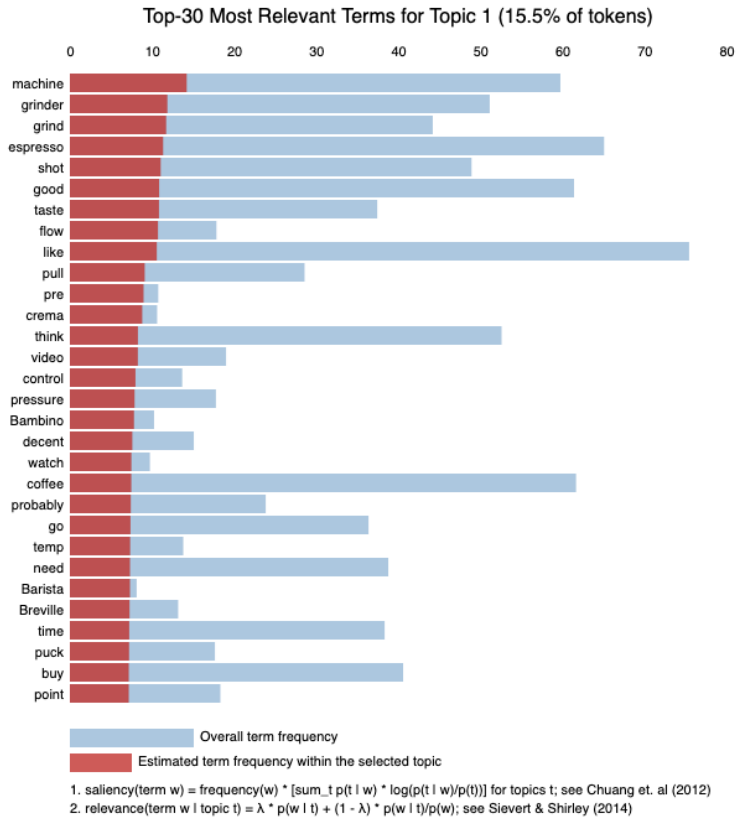
coherence measures over different K

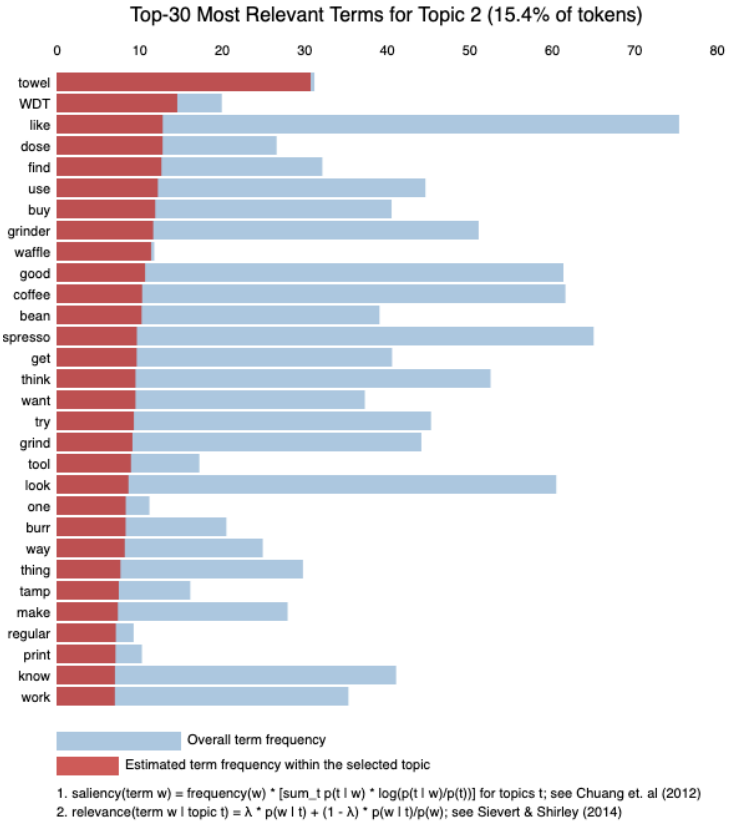


LDA Analysis:

- Coherence score value drops as k increases, indicating that the semantic similarity between the topic words becomes less strong
- This means that having too many topics results in too many words with weak relationships between them
- This has been confirmed during testing for k=20,30 with words = 5,10 for each
- Overall, the best results are with 2-3 topics with 5 words each

- Looking at the **LDA visualization** of the topics, we can see that for Topic 1
    - Words picked up are expected for espresso coffee making machines and taste
    - Machine, grinder, espresso, shot, good, taste, temp, Barista, etc
- For Topics 2,3,4,5 it picks up more granular words for espresso flavour options like
    - Bean, dose, regular, roast, long, nice, sweet, niche, acid, etc
- This aligns with our overall project objective to find ideas for new flavours for coffee
- We can see that the preference range from sweet to regular to acidic, depending on beans and their roast time


Top-30 Most Relevant Terms for Topic 1 (15.5% of tokens)
0
10
20
30
40
50
60
70
80
machine
grinder
grind
espresso
shot
good
taste
flow
like
pull
pre
crema
think
video
control
pressure
Bambino
decent
watch
coffee
probably
go
temp
need
Barista
Breville
time
puck
buy
point
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Top-30 Most Relevant Terms for Topic 2 (15.4% of tokens)
0
10
20
30
40
50
60
70
80
towel
WDT
like
dose
find
use
buy
grinder
waffle
good
coffee
bean
spresso
get
think
want
try
grind
tool
look
one
burr
way
thing
tamp
make
regular
print
know
work
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

wordcloud

- The word combinations show a mix of espresso making machine features along with the type of coffee flavours and how much they are preferred by the customers
- Lot of positive sentiment words like ‘love’, ‘yes’, ‘want’, ‘buy’, etc
- Overall customers are positive about espresso coffee

Sentiment analysis

TextBlob:

- Sentiment(polarity=0.25, subjectivity=0.5785714285714285)
- Polarity indicates the topics are slightly positive with high subjectivity
- Overall indicates that customers making the comments are positive about the espresso features they are looking for and subjective to their individual preferences

vaderSentiment:

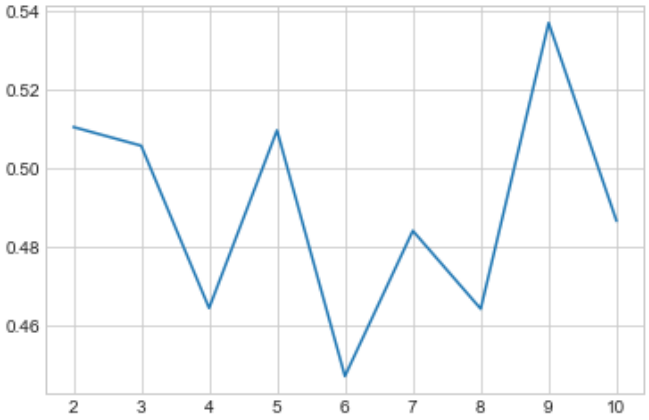
- {'neg': 0.026, 'neu': 0.501, 'pos': 0.473, 'compound': 0.9819}
- Compound score shows that the topic words are very positive and the individual scores too indicate very low negativity and generally neutral or positive comments
- Overall sentiment is positive and further confirms the results of TextBlob sentiment

number of comments: 10523
number of unique words: 15214
total number of words in the corpus: 200470
average number of words in comments: 19.05065095505084
maximum number of words in comments: 477
minimum number of words in comments: 0
median number of words in comments: 11.0
number of unique authors: 3126
number of comments replying to other comments: 6033
number of submissions: 682

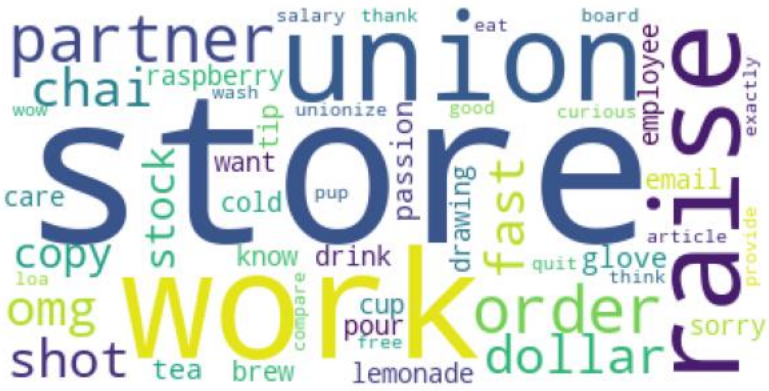
Starbucks
Comments ~ 11,000

Lda coherence = 0.4721744818151736

coherence measures over different K



wordcloud



Sentiment analysis

TextBlob:

- Sentiment(polarity=0.05625000000000001, subjectivity=0.78125)

vaderSentiment:

- {'neg': 0.02, 'neu': 0.605, 'pos': 0.375, 'compound': 0.9702}

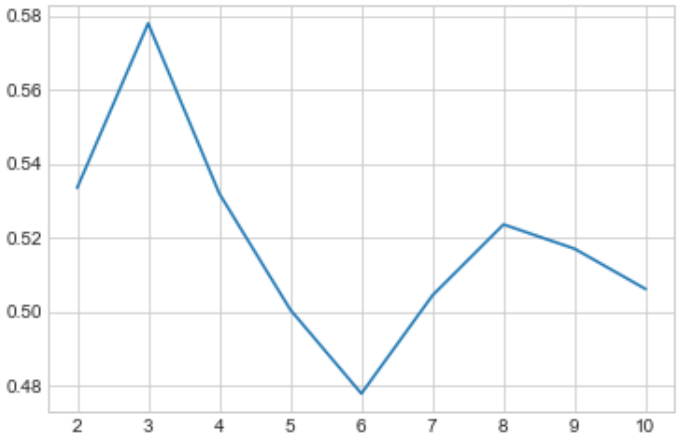
number of comments: 11639
number of unique words: 14209

total number of words in the corpus: 199030
average number of words in comments: 17.100266345906007
maximum number of words in comments: 312
minimum number of words in comments: 0
median number of words in comments: 11.0
number of unique authors: 3605
number of comments replying to other comments: 5810
number of submissions: 858

Coffee
Comments ~ 9,000

Lda coherence = 0.456986178775039

coherence measures over different K



wordcloud

Sentiment analysis

TextBlob:

- Sentiment(polarity=0.17878787878787877, subjectivity=0.543939393939394)

vaderSentiment:

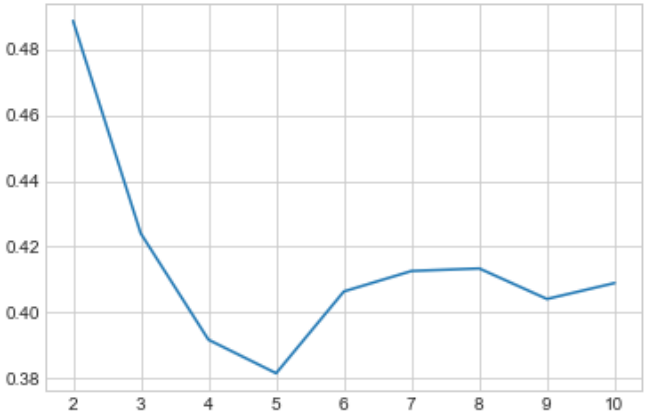
- {'neg': 0.082, 'neu': 0.563, 'pos': 0.355, 'compound': 0.9509}

number of comments: 9175
number of unique words: 15310
total number of words in the corpus: 220392
average number of words in comments: 24.02092643051771
maximum number of words in comments: 659
minimum number of words in comments: 0
median number of words in comments: 15.0
number of unique authors: 3620
number of comments replying to other comments: 5003
number of submissions: 783

Cafe
Comments ~ 1700

Lda coherence = 0.4248070235345594

coherence measures over different K



wordcloud

Sentiment analysis

TextBlob:

- Sentiment(polarity=0.6333333333333333, subjectivity=0.6932098765432099)

vaderSentiment:

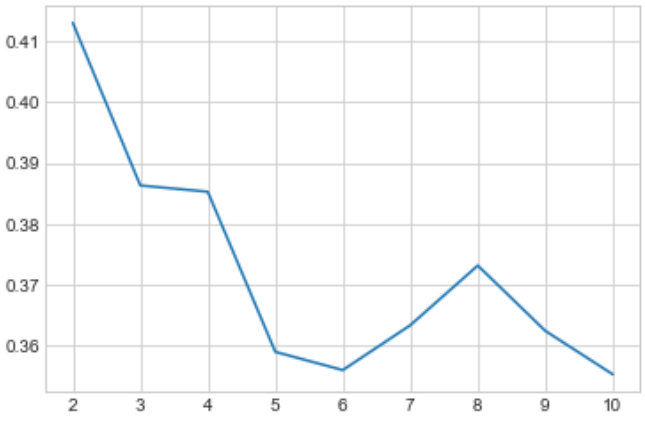
- {'neg': 0.0, 'neu': 0.355, 'pos': 0.645, 'compound': 0.9937}

number of comments: 2326
number of unique words: 7894
total number of words in the corpus: 53253
average number of words in comments: 22.894668959587275
maximum number of words in comments: 400
minimum number of words in comments: 0
median number of words in comments: 13.0
number of unique authors: 851
number of comments replying to other comments: 1204
number of submissions: 271

Coffee_Roaster
Comments ~ 2300

Lda coherence = 0.36698964041292215

coherence measures over different K



wordcloud

Sentiment analysis

TextBlob:

- Sentiment(polarity=0.2313852813852814, subjectivity=0.45)

vaderSentiment:

- {'neg': 0.037, 'neu': 0.614, 'pos': 0.349, 'compound': 0.9628}

number of comments: 1735
number of unique words: 6264
total number of words in the corpus: 35702
average number of words in comments: 20.577521613832854
maximum number of words in comments: 352
minimum number of words in comments: 0
median number of words in comments: 11.0
number of unique authors: 885
number of comments replying to other comments: 813
number of submissions: 220

Meaning of scores:

TextBlob

The `sentiment` property returns a namedtuple of the form `Sentiment(polarity, subjectivity)`. The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.

vaderSentiment

The `compound` score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). This is the most useful metric if you want a single unidimensional measure of sentiment for a given sentence. Calling it a 'normalized, weighted composite score' is accurate.

It is also useful for researchers who would like to set standardized thresholds for classifying sentences as either positive, neutral, or negative. Typical threshold values (used in the literature cited on this page) are:

1. **positive sentiment**: `compound` score >= 0.05
2. **neutral sentiment**: (`compound` score > -0.05) and (`compound` score < 0.05)
3. **negative sentiment**: `compound` score <= -0.05

The `pos`, `neu`, and `neg` scores are *ratios for proportions of text that fall in each category* (so these should all add up to be 1... or close to it with float operation). These are the most useful metrics if you want to analyze the context & presentation of how sentiment is conveyed or embedded in rhetoric for a given sentence.