Assignment 1

R.Walport and R. Raman

For Assignment 1 we tackled yelp data, taking the text from reviews (at this stage not considering other available meta-data, such as the star rating and the users background information) to build our analysis.

The end results have not been altogether successful. Creating topic clusterings via LSI modelling at first derived us very little as food descriptors proved the major defining trait of a restaurant's reviews (words like Thai, noodle, Italian and pasta). As can be seen in the topic clustering below, topic 0 collects the most negative words which is a clear grouping of the most negatively reviewed restaurants but the other topics are dominated by foodstuffs. What's more when we group the restaurants according to which category they most resemble none of our restaurants fall in group zero. More interestingly, and at least showing that the bag of words analysis can recover some information about our restaurants, the topics do give a high precision of restaurant identification. Topic 1 for example pulls out thirteen restaurants all of which are asian (a high precision but low recall since our restaurant sample of 110, contains roughly 50% asian restaurants).

-0.144*"great" + -0.137*"place" + -0.130*"good" + -0.125*"one" + -0.125*"service" + -0.120*"food" + -0.119*"restaurant" + -0.116*"really" + -0.114*"thai" + -0.110*"like"

0.263*"noodle" + 0.234*"thai" + 0.198*"noodles" + -0.169*"italian" + 0.163*"curry" + -0.151*"night" + 0.150*"rice" + 0.149*"like" + 0.148*"fried" + 0.145*"beef"

0.326*"thai" + -0.223*"sum" + -0.213*"dim" + -0.199*"really" + -0.158*"chinese" + 0.126*"salad" + 0.123*"pad" + 0.114*"city" + -0.107*"sure" + -0.105*"far"

0.279*"italian" + 0.172*"dim" + -0.161*"week" + -0.151*"ever" + -0.149*"brunch" + 0.146*"sum" + 0.144*"big" + -0.127*"times" + -0.122*"couple" + 0.116*"old"

-0.244*"like" + -0.236*"brunch" + 0.161*"chinese" + -0.144*"sweet" + -0.128*"curry" + -0.123*"things" + -0.123*"canai" + 0.114*"beef" + -0.113*"amazing" + -0.111*"us"

-0.213*"italian" + -0.180*"last" + 0.166*"really" + 0.162*"salad" + -0.154*"city" + -0.143*"night" + 0.142*"table" + -0.137*"always" + 0.116*"chicken" + 0.115*"lunch"

-0.221*"sum" + -0.214*"dim" + 0.166*"noodles" + 0.157*"noodle" + -0.133*"try" + 0.130*"taiwanese" + 0.129*"flushing" + -0.127*"will" + 0.126*"korean" + 0.119*"trip"

-0.238*"thai" + -0.176*"bbq" + 0.168*"old" + -0.161*"korean" + 0.145*"japanese" + -0.139*"ordered" + 0.116*"get" + -0.111*"dim" + -0.110*"sum" + 0.110*"paella"

-0.218*"italian" + 0.157*"found" + -0.156*"less" + -0.144*"one" + -0.142*"favorite" + 0.141*"thai" + 0.137*"night" + -0.137*"bbq" + -0.124*"noodle" + -0.119*"great"

0.168*"sangria" + -0.158*"one" + 0.145*"week" + -0.144*"get" + -0.138*"japanese" + 0.136*"brunch" + 0.132*"experience" + 0.131*"bread" + 0.129*"bean" + 0.127*"overall"

Overall though this wasn't really what we were looking for. So we made the decision to create a second stop word list of food terms, we were looking for restaurant quality and not type so food nouns are not useful differentiators and ran the same analysis again.

0.159*"great" + 0.145*"good" + 0.142*"place" + 0.135*"one" + 0.135*"service" + 0.134*"like" + 0.124*"really" + 0.116*"best" + 0.111*"go" + 0.108*"restaurants"

0.347*"like" + -0.193*"night" + 0.170*"pretty" + 0.168*"ever" + 0.161*"sweet" + 0.125*"things" + 0.120*"nyonya" + -0.116*"table" + -0.115*"city" + 0.113*"highlights"

-0.176*"really" + -0.171*"us" + -0.139*"like" + -0.133*"say" + 0.131*"little" + 0.130*"since" + -0.125*"made" + 0.117*"special" + 0.115*"week" + -0.115*"old"

-0.195*"pretty" + -0.181*"old" + 0.144*"thought" + 0.121*"friends" + 0.117*"know" + 0.116*"reviews" + -0.113*"attentive" + -0.112*"several" + -0.111*"staff" + 0.110*"one"

-0.159*"years" + 0.157*"brunch" + -0.156*"last" + -0.148*"traditional" + -0.136*"authentic" + 0.136*"great" + -0.135*"old" + 0.133*"everything" + 0.133*"atmosphere" + 0.126*"amazing"

0.179*"brunch" + 0.160*"city" + -0.144*"really" + 0.140*"best" + 0.137*"last" + 0.134*"found" + 0.133*"favorite" + -0.128*"less" + -0.125*"lunch" + 0.120*"night"

-0.217*"spicy" + -0.173*"ordered" + -0.168*"fried" + -0.152*"green" + 0.137*"will" + -0.135*"walked" + -0.132*"just" + 0.130*"nice" + -0.129*"took" + -0.123*"dumplings"

-0.232*"week" + -0.208*"really" + -0.195*"experience" + -0.193*"brunch" + -0.166*"overall" + -0.150*"bean" + -0.147*"went" + 0.138*"spicy" + -0.127*"bread" + 0.123*"one"

-0.205*"one" + -0.173*"favorite" + -0.148*"big" + -0.139*"always" + -0.129*"staff" + -0.127*"less" + 0.119*"though" + -0.116*"city" + 0.115*"something" + -0.111*"fried"

0.156*"order" + 0.147*"times" + 0.135*"terrible" + -0.130*"ambiance" + 0.127*"never" + 0.121*"get" + -0.120*"evening" + 0.111*"rude" + -0.108*"tasty" + 0.106*"now"

These results show a good deal more promise. Topic 0 groups words that are almost exclusively positive (notwithstanding negation references which we can't do much about). Many of the other categories are less clear though Topic 9 captures many of the most negative words. What we found here is that 85 of the restaurants fell most into Topic 0. A mere 5 restaurants fell into bucket 9 so we looked closer at these. Unfortunately when correlated to the restaurant's cleanliness ratings and overall violation count and type, our results to do not hold up at all (indeed all five hold A ratings, with close to zero violations over the last five years...).

There are two possible reasons for this as far as we can see. Firstly, there really is no correlation between restaurant health code violations and yelp reviews (which is not completely implausible) but more likely to our minds, it is because we have acquired insufficient raw data.

To add another check for this we ran LDA topic modelling to see if the results were any better:

---

topic #0: 0.008*amazing + 0.007*brunch + 0.007*excellent + 0.006*price + 0.006*theater + 0.006*friends + 0.006*great + 0.006*said + 0.006*two + 0.006*different

topic #1: 0.010*like + 0.008*years + 0.008*get + 0.007*manhattan + 0.007*go + 0.007*reviews + 0.006*wait + 0.006*place + 0.006*really + 0.006*good

topic #2: 0.009*less + 0.007*tourist + 0.007*can + 0.006*worth + 0.006*try + 0.006*group + 0.006*going + 0.006*filling + 0.006*appetizer + 0.006*large

topic #3: 0.010*mexican + 0.008*vegetarian + 0.007*great + 0.007*one + 0.007*sweet + 0.006*enough + 0.006*upscale + 0.006*dish + 0.006*make + 0.006*bad

topic #4: 0.009*best + 0.008*definitely + 0.007*thing + 0.007*sticky + 0.007*ever + 0.007*fried + 0.006*flushing + 0.006*pricey + 0.006*new + 0.006*good

topic #5: 0.007*big + 0.006*inside + 0.006*terrible + 0.006*many + 0.006*great + 0.006*side + 0.006*quite + 0.006*friendly + 0.005*west + 0.005*ny

topic #6: 0.009*spicy + 0.008*family + 0.008*cuisine + 0.008*evening + 0.007*mom + 0.007*authentic + 0.007*much + 0.007*4 + 0.007*expensive + 0.006*peking

topic #7: 0.008*quite + 0.008*ordered + 0.008*table + 0.008*wanted + 0.008*green + 0.008*really + 0.007*average + 0.007*fun + 0.007*bit + 0.006*craving

topic #8: 0.007*servers + 0.007*dishes + 0.007*know + 0.007*real + 0.006*summer + 0.006*street + 0.006*times + 0.006*delicious + 0.006*bread + 0.006*date

topic #9: 0.009*went + 0.009*pretty + 0.008*last + 0.007*city + 0.007*night + 0.007*better + 0.007*service + 0.007*everything + 0.007*come + 0.007*since

---

These clusterings were no more helpful with few clear categories emerging. Increasingly we began to believe we simply lacked enough data.

One of the biggest problems of our analysis was the relative lack of reviews. Though Yelp has a highly usable API it returns only a selection of the most recent reviews and only a stub of each review (which it's own internal algorithm selections). This has introduced a far greater error margin as with the reduced numbers, the impact of one grumpy customer could potentially be far too greatly weighted. For example in the case of a restaurant called "El Quijote", we happened to grab a series of negative reviews. Though it's overall score of 3.5 stars on yelp is fairly low, these very negative scores were not a true reflection of its aggregate reviews.

To solve this I believe we need to go back and tackle Yelp as a website as a scraping exercise rather than using the API. We haven't had a chance to do this for this assignment unfortunately (we really discovered too late in the day that our data was insufficient).