**Q- Explain two alternative methods to expand this lexicon automatically. Which of these methods should result in the larger lexicon and why?**

A- Determine the emotion of these subjective word/phrases.

Dictionary-based: find synonyms/antonyms of seed emotion words in dictionaries like WordNet

Corpus-based: find synonyms/antonyms of seed emotion words in corpora

Semi-automatically created resources, such as:

SentiWordNet: Wordnet is a database with words grouped into sets of synonyms (synsets), and organised by semantic relations between them: synonyms, antonyms, hypernyms, etc. SentiWordNet is a version of it with one of three sentiment scores for each synset: positivity, negativity, objectivity.

Semi-automatically created from seed words: start with seed positive and negative words:

• Search for synonyms/antonyms in dictionaries like WordNet; OR

• Build patterns from seed words/phrases to search on large corpora, like the Web: ⇧ “beautiful and” (+) ⇧ “low cost but” (-) ⇧ “very nice and ” (+)

**Q-According to Bing Liu’s model, an opinion is said to be a quintuple (oj, fjk, soijkl, hi, tl). Explain each of these elements and exemplify them with respect to the following text. Identify the features present in the text, and for each indicate its sentiment value as either positive or negative. Discuss two language processing challenges in automating the identification of such elements.**

A- An opinion is a quintuple (o, f, so, h, t), where:

• o is a target object.

• f is a feature of the object oj.

• so is the sentiment value of the opinion of the

• opinion holder h (usually the author of the post)

• on feature f of object o at time t.

soijkl is positive, negative, neutral, or a more granular rating, such as 1-5 stars as in movie reviews

**“I have just bought the new iPhone 12. It is a bit heavier than the iPhone 11, but it is much faster. The camera lenses are also much better, taking higher resolution pictures. The only big disadvantage is the cost: it is the most expensive phone in the market. Michael Jordan, 12/08/2020.”**

o- iPhone 12

f- weight, performance, camera, price

so- negative, positive, positive, negative

h- Michael Jordan

t- 12/08/2020

**Q- Explain the intuition behind using a Naive Bayes classifier for Sentiment Analysis. Give the general classifier equation as part of your answer. What are the main components in this classifier? Give two types of features that could be used and provide examples for these types of features.**

A- @ Is this a good solution? Is it robust? — It’s simple and will work well if data is not sparse

@ What is the role of the prior? — Prior is very important esp. on biased cases

© How can we improve this solution? © Other features? Are we missing out critical information?

— Using all words (in Naive Bayes) works well in some tasks

— Finding subsets of words may help in other tasks

— Using only adjectives can be limiting. Verbs like hate, dislike; nouns like love; words for inversion like not; intensifiers like very

— Pre-built polarity lexicons can be helpful

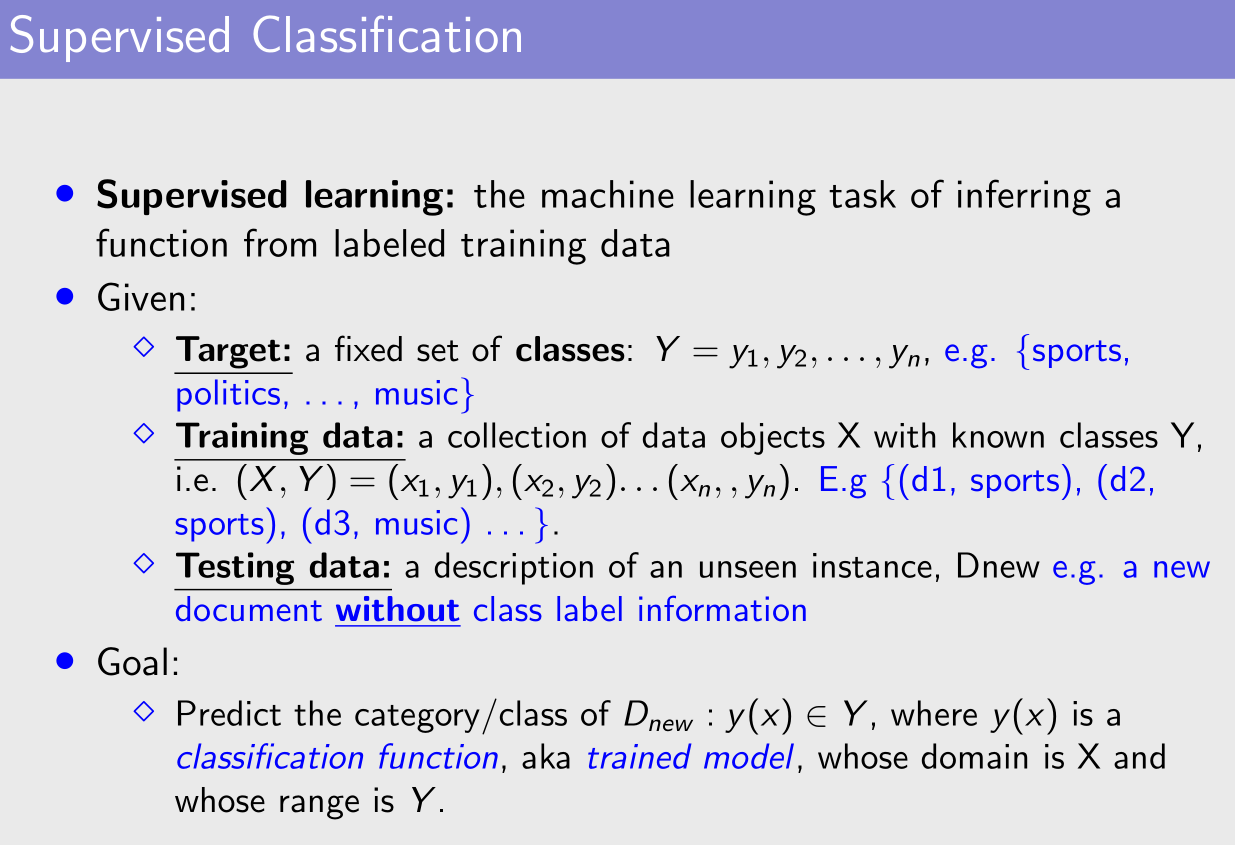
— Negation is important

© Other algorithms? — MaxEnt & SVM tend to do better than Naive Bayes

@ What about non-binary classification (e.g. 5-grades of sentiment)? — 5-class ordinal classification or regression algorithms can be used

**Q- Explain the corpus-based supervised learning approach to Sentiment Analysis in general terms, i.e. in terms of inputs, outputs and processes involved.**

A-



**Q- Explain the concept “independence assumption” used by a Naive Bayesian Classifier.**

A- Assume A and B are Boolean Random variables. Then “A and B are independent” if and only if

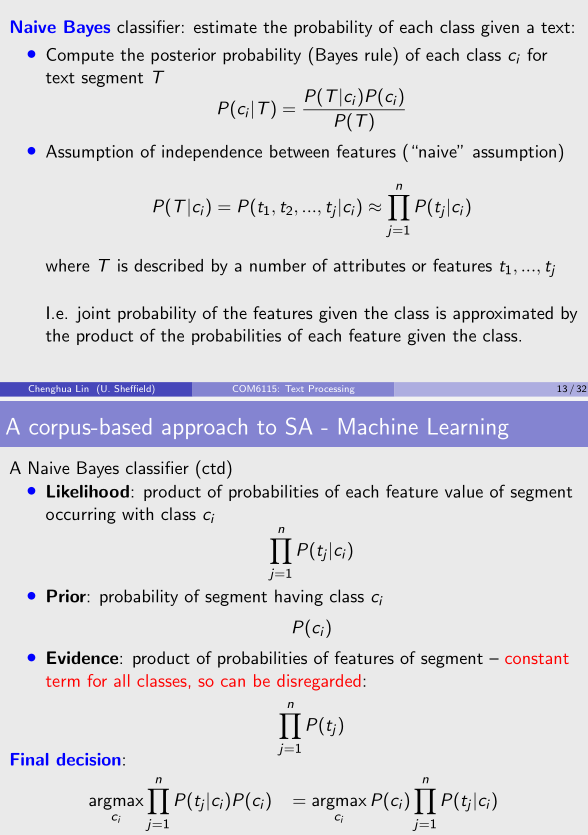
P(A|B) = P(A)

Features (term presence) are independent of each other given the class:

P(X1, ..., X5|Y ) = P(X1|Y ) • P(X2|Y ) • ... • P(X5|Y )

**Q- Explain how a Naive Bayes classifier can be trained and then used to predict the polarity class (positive or negative) of a subjective text. Be sure to give the mathematical formulation of the Naive Bayes classifier.**

A-



**Q- What are the stages of processing commonly followed within natural language generation (NLG) systems? For each stage, please explain its purposes.**

A- Usually three stages, Not including data analysis

• Document planning: decide on content and structure of text

Document Planning- Content selection: of the zillions of things I could say, which should I say?

⇧ Depends on what is important, ⇧ What makes good narrative, ⇧ What is easy to say

-Structure: How should I organise this content as a text?

⇧ What order do I say things in?, ⇧ Rhetorical structure?

• Microplanning: decide how to linguistically express text (which words, sentences, etc to use)

Microplanning

- Lexical/syntactic choice: Which words and linguistic structures to use?

- Aggregation: How should information be distributed across sentences and paras

- Reference: How should the text refer to objects and entities?

• Realisation: grammatical details E.g. children vs. childs, an apple vs. a apple

-Grammars (linguistic): Form legal English sentences based on decisions made in previous stages

i) Obey sub-languages, i.e., language of a restricted domain, particularly a technical domain.

ii) genre constraints, e.g., scientific writing vs. social media text, etc

-Structure: Form legal HTML, RTF, or whatever output format is desired.

**Q-An NLG system needs to take care of details of language such as morphological details. How does inflectional morphology differ from derivational morphology? Explain with examples from the English language.**

A- Morphology

In linguistics, morphology is the study of words, how they are formed, and their relationship to other words in the same language. E.g.,- Variations of a root form of a word, e.g., prefixes, suffixes

• Inflectional morphology - same core meaning

⋄ plurals, past tense, superlatives, e.g., dog, dogs

⋄ part of speech unchanged

• Derivational morphology - change meaning

⋄ prefix re means do again: reheat, resit

⋄ suffix er means one who: teacher, baker

⋄ part of speech changed

**Q-Explain three metrics to evaluate the quality of binary (negative/positive) sentiment analysis systems. Give their intuitions and show their formulae.**

**Q- Define the precision and recall measures in Information Retrieval.**

A- Principal metrics — borrowed from information retrieval — are:

© Precision (how much of what system returns is correct)

© Recall (how much of what is correct system returns)

© F-measure (a weighted combination of precision and recall)

**Q-What is the first step to detect the sentiment in these two sentences?**

A- Subjectivity analysis

Subjectivity classification is often the first step for sentiment analysis: subjective versus objective texts, e.g.:

® Objective: bought an iPhone a few days ago.

- Subjective: It is such a nice phone.

However: Subjective sentences do not always express positive or negative opinions, e.g.: I think he came yesterday.

® Objective sentences can express opinion indirectly, e.g.: My phone broke in the second day

**Q-Explain what is meant by an geverfed index and why such indices are important in the context of Information Retrieval.**

A- A basic inverted file index

©records for each term, the ids of the documents in which it appears

© only matters if it does or does not appear — not how many times

A more sophisticated version ...

© also record count of occurrences within each document

© help find documents more relevant to query

A more sophisticated version . . .

⋄ also record position of each term occurrence within documents

⋄ may be useful for searching for phrases in documents

**Q-Discuss the advantages and disadvantages of Boolean versus ranked approaches to Information Retrieval.**

A-The Boolean model: summary

i) Documents either match or don’t match

©Expert knowledge needed to create high-precision queries —> OK for expert users

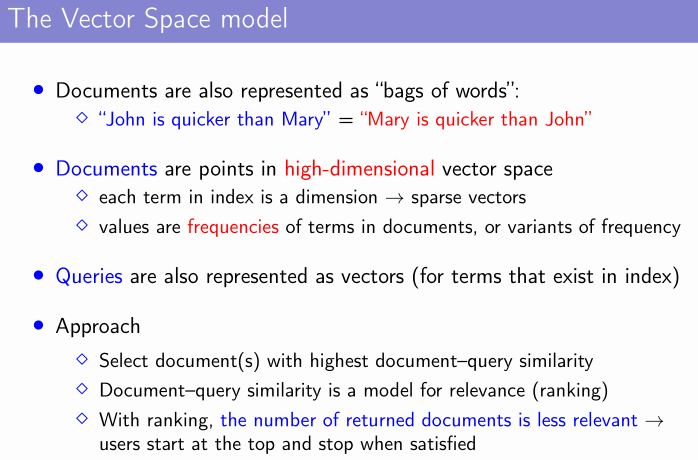
© Often used by bibliographic search engines (library)

ii) Not good for the majority of users

© Most users not familiar with writing Boolean queries —> not natural

© Most users don't want to wade through lists of 1000s unranked results — unless very specific search in small collections

© This is particularly true of web search — large set of doc



Past Paper

2013-2014

**Section A**

**a) In the context of Information Retrieval, explain the difference between algorithms that perform boolean search and algorithms that perform a ranked search. What type of algorithm would be better for a regular user (such as an undergraduate student in the Humanities area) who is using a search query with multiple terms, which he/she expects to appear in many documents? Explain the reasons behind your choice of algorithm type.**

Answer:

​ For the boolean search model, the model just needs to decide the whether the document is relevant or not. And the presence of the terms is very essential and suffient for the search. For the operator, the boolean search model exploits the boolean operators like AND and OR. The boolean query provides the logical result for deciding whether the document should be returned.

​ For the ranked search method like the vector space model, it relys on the frequency of terms and maybe some weight of terms could affect the result of the search. The documents are point in high-dimension vector space and its value is the frequency of the terms. Also, the query should be represented as the vectors. This method record the vectors for each documents and queries, finally calculate the similarity metrics which can be interpreted as the normalised correlation coefficient.

​ For the choice of the searching algorithm, the ranked algorithm should be chosen. Because the boolean algorithm is difficult and unnatural for the newbies. And the users don’t want the plenty of ranked results list. For the ranked algorithm, it’s very easy to generate the result lists and pick up the relevant documents.

**\*\*d)Assume we have a small set of seed words with positive and negative opinions, e.g.: positive = {good, fast, cheap} and negative = {slow, boring, fragile}. Explain the two most common (semi-)automated approaches to expand these sets with more opinion words or phrases to create lexica for Sentiment Analysis, providing examples whenever possible. Give one advantage and one disadvantage of each approach. \*\***

Answer:

​ This model is created from the seed words which comes from the supervised seed positive and negative words. It can be divided into two parts ways, Dictionary-based and Corpus-based approach. For the Dictionary-based, the users should find the synonyms/antonyms of seed emotion words in dictionaries like WordNet. Otherwise, users can build the patterns from the seed words/phrases to search on large corpora. For the Corpus-based approach, the examples are annotated with sentiment are used withe machine learning algorithms to learn a classifier for each sentence and document. The users can rely on the manully way (gold-standards) and crow-annotated resources like Amazon Product Resource. This approach can be divided into the two steps: subjectivity classifier : first run binary classifier to identify and then eliminate objective segments. Subjectivity Classifier with remaining segments: learn how to combine and weight different attributes to make predictions.

​ For the advantages for the Dictionary-based approach, it doesn’t need the labelled data and the procedure of learning is not required. For disadvantages, it requires powerful linguistic resources which is not always available, so users cannot guarantee the efficiency of this approach.

​ For the advantages for the Corpus-based approach, it really easy to implement anf usually works well. As for the disvantage, the assumption of this approach is hard to estimate.

**Section B**

2.In the context of Information Retrieval, given the following documents:

Document 1: Sea shell, buy my sea shell! Document 2: You may buy lovely SEA SHELL at the sea produce market. Document 3: Product marketing in the Shelly sea is an expensive market. and the query: Query 1: sea shell produce market

**a)Apply the following term manipulations on document terms: stoplist removal, capitalisation and stemming, showing the transformed documents. Explain each of these manipulations. Provide the stoplist used, making sure it includes punctuation, but no content words.**

Answer:

​ For the stop removal list, it should exclude the “non-content” words, so the stop list should be “my, may ,at, you ,the, in, is, an, , , ! , .”.

​ For the capitalisation, turn all the words to lower case.

​ For the stemming, turn the marketing to market. product to produce

​ So the transformed document should be:

Document 1: sea shell buy sea shell

Document 2: buy lovely sea shell sea produce market

Document 3: produce market shelly sea expensive market

**b)Show how Document 1, Document 2 and Document 3 would be represented using an inverted index which includes term frequency information.**

Answer:

​ sea: (1,1) (1,4) (2,3) (2,5) (3,4)

​ shell: (1,2) (1,5) (2,4)

​ buy: (1,3) (2,1)

​ lovely:(2,2)

​ produce: (2,6) (3,1)

​ market: (2,7) (3,2) (3,6)

​ shelly:(3,3)

​ expensive: (3,5)

**c)Using term frequency (TF) to weight terms, represent the documents and query as vectors. Produce rankings of Document 1, Document 2 and Document 3 according to their relevance to Query 1 using two metrics: Cosine Similarity and Euclidean Distance. Show which document is ranked first according to each of these metrics.**

Answer:

​ **TF:**

​ Query: sea(1) shell(1) produce(1) market(1)

​ Document1: sea(2) shell(2) produce(0) market(0) buy(1)

​ Document2: sea(2) shell(1) produce(1) market(1) buy(1) lovely(1)

​ Document3: sea(1) shell(0) produce(1) market(2) expensive(1)

​ **Evaluation:**

​ Similarity cosine:

​ Document1: (2×1+2×1+0×1+0×1+0×1) / (2× √(4+4+0+0+1)) = 4/6

​ Document2: (2×1+1×1+1×1+1×1+0×1+0×1) / (2×√(4+1+1+1+1+1)) = 5/6

​ Document3: (1×1+0×1+1×1+2×1+0×1) / (2×√(1+0+1+4+1) = 4/5.29

​ Euclidean Distance：

​ Document1: √(1+1+1+1+1) = √5

​ Document2: √(1+0 +0 + 0 + 1 + 1) = √3

​ Document3: √(0+1+0+1+1) = √3

​ For the best similarity, we should choose the lowest value document.

**d)Explain the intuition behind using TF.IDF (term frequency inverse document frequency) to weight terms in documents. Include the formula (or formulae) for computing TF.IDF values as part of your answer. For the ranking in the previous question using cosine similarity, discuss whether and how using TF.IDF to weight terms instead of TF only would change the results.**

Answer:

​ For using the Terms frequency, the result of the model may be affected by the freqency of the terms in documents, so we should consider the less common terms could be more useful to find the relevant documents, so we use the inverse document frequency to avoid this situation happening.

​ we need the document frequency (df) which include the key terms from the query.

​ sea : idf:0 log1 = 0

​ shell: idf: log(3/2)=0.17609125905

​ buy: idf: log(3/2)=0.17609125905

​ lovely: idf: log(3/1)=0.477

​ produce: idf: log(3/2)=0.17609125905

​ market: idf: log(3/2)=0.17609125905

​ shelly: idf: log(3/1) = 0.477

​ expensive: idf: log(3/1)=0.477

​ **Evaluation**:

| **Terms** | **Query** | **Document1** | **Document2** | **Document3** |
| --- | --- | --- | --- | --- |
| Sea | 0×1/4 | 0×2/5 | 0×2/6 | 0×1/5 |
| Shell | 0.18×1/4 | 0.18×2/5 | 0.18×1/6 | 0.18×0 |
| Buy | 0 | 0.18×1/5 | 0.18×1/6 | 0.18×0 |
| Lovely | 0 | 0.477×0 | 0.477×1/6 | 0.477×0 |
| Produce | 0.176×1/4 | 0.176×0 | 0.176×1/6 | 0.176×1/5 |
| Market | 0.176×1/4 | 0 | 0.176×1/6 | 0.176×1/5 |
| Shelly | 0 | 0 | 0 | 0.477×1/5 |
| Expensive | 0 | 0 | 0 | 0.477×1/5 |
|  | 0.20925 | 0.108 | 0.1975 | 0.0892 |

​ Document1: (0+(0.045×0.072)) / 0.20925×0.108

​ Document2: (0+(0.03×0.045)+(0.044×0.0293)×2) / 0.20925×0.1975

​ Document3: (0+(0.044×0.0352)) / 0.20925×0.0892

**e)Explain the metrics Precision, Recall and F-measure in the context of evaluation in Information Retrieval against a gold-standard set, assuming a boolean retrieval model. Discuss why it is not feasible to compute recall in the context of searches performed on very large collections of documents, such as the Web.**

Answer:

​ **Recall**： the proportion of the relevant documents

​ **Precision**: the proportion of retrieved documents that are relevant

​ Precision and Recall address the relation between the retrieved and relevant sets of documents.

​ **F-measure**: combines precision and recall into a single figure,gives equal weight to both. It is the F is a harmonic mean which penalises low performance in one value more than arithmethic mean.

​ Because the there are tremendous web pages which is included in the Internet, the mount of the retrieved pages is pretty small because of the assumption and limitation of algorithms.

4.

**a) Differentiate subjectivity from sentiment. How are the tasks of Subjectivity Classification and Sentiment Analysis related?**

Answer:

​ As for the rule-based subjectivity classifier, the task is to search for the emotion words lexicon and determine the sentence/document is objective or subjective.

​ As for the rule-based sentiment classifier, the task is to determine the document/sentence shows the positive sentiment or negative sentiment by counting the value of lexicons and build the judgement model.

**b)Explain the steps involved in the lexicon-based approach to Sentiment Analysis of features in a sentence (e.g. features of a product, such as the battery of a mobile phone). Discuss the limitations of this approach.**

Answer:

​ As for the binary approach, the input is the sentences s and product features f, the output is the attitude of this feature is positive,negative or netural.

​ Step1: the model should pick up all the sentences which contains the fratures and lexicons about the attitude.​ Step2: the model should select all the emotion words in sentence​ Step3: assign the values of emotion words, 1=positive, 0=netural, -1=negative​ Step4: sum up the orientation and assign the orientation to (f,s)

​ As for the shortcoming for intensifiers, the gradable approach assign the different levels to the emotional content. The process is similar to the binary approach, the final decision is based on the all emotion words.

​ The disadvantage is requiring a lexicon of emotion words which should be fairly comprehensive,covering new words.abbreviations,misspelled words.

**c)Explain the graded lexicon-based approach for Sentiment Analysis.Given the following sentences and opinion lexicon (adjectives only), apply the weighted lexical-based approach to classify EACH sentence as positive, negative or objective. Show the final emotion score for each sentence. In addition to use of the lexicon, make sure you consider any general rules that have an impact in the final decision. Explain these rules when they are applied.**

**Lexicon:** **boring -3** **brilliant 2** **good 3** **horrible -5** **happy 5**

Graded lexicon-based Approach:

​

(S1)He is brilliant but boring.\*\*

Diminisher rule: the weight should be substracted from the positive terms.

emotion(brilliant) = 2 emotion(boring) = -3 , so the decision value is 2-3 = -1

**(S2) I am not good today.**

Negation rule: when the neighbourhood area occurs the negation words, the value should be decreased by 1 and inverted.

emotion(good) = 3, the decision value is -(3-1) = -2

**(S3) I am feeling HORRIBLE today, despite being happy with my achievement.**

Capitalization rule : the value of emtions words should be increased by 1 for positive words,-1 for negative words. Diminisher rule: the weight should be substracted from the positive terms.

emotion(horrible) = -5, capitalization rules -> emotion(horrible) = -6 emotion(happy) = 5

Decision value = -1

**(S4) He is extremely brilliant but boring, boring.**

Intensifier rule: the weight of extremely is 2, so the emotion(brilliant) = 4, emotion(boring) = -3, so the decision value is -2.

**d)Specify the five elements of Bing Liu’s model for Sentiment Analysis, and exemplify them with respect to the following text. Identify the features present in the text, and for each indicate its sentiment value as either positive or negative. Discuss two language processing challenges in automating the identification of such elements.**

“I am in love with my new Toshiba Portege z830-11j. With its i7 core processors, it is extremely fast. It is the lightest laptop I have ever had, weighting only 1 Kg. The SSD disk makes reading/writing operations very efficient. It is also very silent, the fan is hardly ever used. The only downside is the price: it is more expensive than any Mac. Lucia Specia, 10/04/2012.”

Answer:

​ As for the five elements, they should be:

e\_i : the name of the entity

a\_ij: an aspect of the e\_i

oo\_ijkl: the orientation of the opinion about the aspecrt a\_ij of the entity.

h\_k: the opinion holder

t\_l: the time when opinion is expressedx by h\_k

​ The name of the entity is Toshiba Portege z830-11j. For the many aspects, they should be i7 core processors,SSD disk,weighting,fan,price.The orientation of each opinion about the aspect, for the general idea about this product, the holders said love, it’s the positive. For the aspect i7 core processors, the holder said fast, it is positive. For the weight, the holder said it’sthe lightest, the orientation is positive. As for the SSD disk, the holder said efficient and silent, so the positive. For the fan, it said hardly ever used, it’s netural. The aspect privr, the more expensive, so it’s negative. Aslo, you can generate the quintuple to show the result, like below:

​ (Toshiba Portege z830-11j, i7 core processors, positive,Lucia Specia,10/04/2012)

​ Challenge: 1) Maybe it’s very difficult to comfirm the entity name because of variations and abbreviation. 2) It’s very difficult to distinguish the subjective and objective sentences because of the language.

**e) Differentiate direct from comparative Sentiment Analysis. What are the elements necessary in comparative models of Sentiment Analysis?**

The comparative SA can contrast direct opinion verusu more comparativve opinions such as direct sentiment expressions on target objects and comparisons expressing similarities or differences.

2016-2017

**Section A**

**a)Explain briefly the intuition behind the PageRank algorithm. Discuss how it can documents that are ranked equally “relevant” according to the similarity score given by the vector space model.**

PageRank is designed to use the structure of a graph to quantify the importance of each node in that graph. Accordingly, every usage of PageRank outside of a web context must maintain some notion of importance, even if the interpretation of the importance of a node varies from application to application.

PageRank can be thought of as a fluid that circulates throughout a network, passing from node to node across edges, and pooling at the nodes that are the most important

Intuition: PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.