Text processing

问题:SA:approaches ppt 1.at least 2 不清楚不明白怎么回事，

2.确定如何分辨objective and subjective

**SA**

2013-2014

a) Differentiate subjectivity from sentiment. How are the tasks of Subjectivity Classification and Sentiment Analysis related? [10%]

**Answer:**Subjectivity classification is often the first step for sentiment analysis

Subjective sentences do not always express positive or negative opinions

Objective sentences can express opinion indirectly

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. [20%]

**Answer:**(1) Identify and extract object features that have been commented on by an opinion holder (e.g., a reviewer).

(2) Determine whether the opinions on the features are positive, negative or neutral.

(3) Group synonym features, e.g. screen and touch screen.

(4) Optional: produce a feature-based opinion summary of multiple reviews.

Limitations: documents and sentences may contain mixed opinions and analysis at this level does not identify specifically what people like/dislike.

c) 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. [20%]

Lexicon: boring -3

brilliant 2

good 3

horrible -5

happy 5

(S1) He is brilliant but boring.

(S2) I am not good today.

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

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

**Answer:** (S1) 2+(-3) =-1 negative

(S2) 3-1=2 2=>-2 negative

Negative rule: “not” is detected in neighborhood (of 5 words around);

so emotional valence of word is decreased by 1 and sign is inverted

(S3) -5-1=-6 -6+5=-1 negative

Capitalization rule: One word is capital writing.

Add +1 to positive words.

Add -1 to negative words.

(S4)1+2=3 3+(-3) +(-3) =-2 negative

Intensifier rule: Needs a list of intensifiers: “very”, “definitely”, “extremely”, etc.

The weight is added to positive terms.

The weight is subtracted from negative terms

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. [30%]

“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:** A quintuple (oj, fjk ,soijkl, hi ,tl):

oj is a target object

fjk is a feature of the object oj

soijkl is the sentiment value of the opinion of the opinion holder hi (usually the author of the post) on feature fjk of object oj at time tl .

oj: Toshiba Portege z830-11j

fjk: Toshiba Portege z830-11j ,i7 core processors, SSD disk, fan, price

soijkl: positive, positive, positive, positive, negative

opinion holder hj: I, I, I, I, I

time tl: 10/04/2012

Identifying target objects

1.Named Entity Recognition: well-known tools based on gazetteers and simple context rules.

Need good gazetteers: Web is dynamic, new products appearing everyday.

Standard NE recognisers will not work for objects like names of movies

2.Bootstrap from seed gazetteers: e.g. if know that iPhone 4 is an object, can find out that iPhone 5 is also an object

Co-reference (and synonym) resolution

1. Important to resolve objects and features

e) Differentiate direct from comparative Sentiment Analysis. What are the elements necessary in comparative models of Sentiment Analysis? [20%]

**Answer:** 1. Gradable Non-equal gradable: Relations of the type greater or less than. E.g.:

“lenses of camera A are better than those of camera B”

2. Equative: Relations of the type equal to. E.g.: “camera A and camera B both come in 7MP”

3. Superlative: Relations of the type greater or less than all others. E.g.: “camera A is the cheapest camera available in market”

4. Non-gradable comparisons: Relations that compare aspects of two or more entities, but do not grade them. e.g., “Coke tastes differently from Pepsi.”

2014-2015

1. Consider the two sentences:

• My new phone works well, is very pretty and much faster than the old one.

• My new phone has 32GB of memory and plays videos.

What is the first step to detect the sentiment in these two sentences? Should both these sentences be addressed in the same way by Sentiment Analysis approaches? If not, explain a common approach to select only relevant sentences for Sentiment Analysis. [20%]

**Answer:** Subjectivity classification is often the first step for sentiment analysis.

No, one is objective sentence and another is subjective sentence.

Choose the first one. It uses one type of the comparative relations called gradable. And pretty is a positive word. Also, it uses much faster to show the speed of the phone. Thus, the sentiment analysis is positive.

1. Given the following sentences S1 to S4 and opinion lexicon of adjectives, apply the weighted lexical-based approach to classify EACH sentence as positive, negative or objective. Show the final emotion score for each sentence, and also how it was generated. In addition to using the lexicon, make sure you consider any general rules that have an impact on the final decision. Explain these rules when they are applied. [15%]

Lexicon: awesome 5

boring -3

brilliant 2

funny 3

happy 4

horrible -5

(S1) He is brilliant and funny.

(S2) I am not happy with this outcome.

(S3) I am feeling AWESOME today, despite the horrible comments from my supervisor.

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

**Answer:** (S1) 2+3=5 positive

(S2) 4-1=3 3=>-3 negative

Negative rule: “not” is detected in neighborhood (of 5 words around);

so emotional valence of word is decreased by 1 and sign is inverted

(S3) 5+1=6 6+(-5) =1 positive

Capitalization rule: One word is capital writing.

Add +1 to positive words.

Add -1 to negative words.

(S4) 2+1=3 3+(-3) +(-3) =-3 -3-1=-4 -3+(-4)=-7

Intensifier rule: Needs a list of intensifiers: “very”, “definitely”, “extremely”, etc.

The weight is added to positive terms.

The weight is subtracted from negative terms

1. 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. [25%]

“I have just bought the new iPhone 5. It is a bit heavier than the iPhone 4, 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. Lucia Specia, 12/08/2014.”

**Answer:** oj is a target object

fjk is a feature of the object oj

soijkl is the sentiment value of the opinion of the opinion holder hi (usually the author of the post) on feature fjk of object oj at time tl .

oj: iPhone 5

fjk: iPhone 5, the camera lenses, price

soijkl: neutral, positive, positive, negative(明天问)

opinion holder hj: I, I, I, I

time tl: 12/08/2014

Identifying target objects

1.Named Entity Recognition: well-known tools based on gazetteers and simple context rules.

Need good gazetteers: Web is dynamic, new products appearing everyday.

Standard NE recognisers will not work for objects like names of movies

2.Bootstrap from seed gazetteers: e.g. if know that iPhone 4 is an object, can find out that iPhone 5 is also an object

Co-reference (and synonym) resolution

1.Important to resolve objects and features

1. Assume a lexicon-based approach to binary Sentiment Analysis. A manually created initial lexicon is available which contains only three positive words:

• good

• nice

• excellent

and three negative words:

• bad

• terrible

• poor

This lexicon needs to be expanded in order for the approach to be effective in a realistic task. Explain two alternative methods to expand this lexicon automatically. Which of these methods should result in the larger lexicon and why? [20%]

**Answer:** 1. Dictionary-based: find synonyms/antonyms of seed emotion words in dictionaries like WordNet

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

Dictionary-bases usually focus on one field such LIWC for psychologists. And corpus-bases can extend by machine learning to learn a classifier for each sentence/document. Corpus is a collection of text segments (e.g. webpages, blog posts, reviews, tweets, etc) with humanly-annotated emotional indicators (e.g. positive, negative, etc).

e) 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. [20%]

**Answer:贝叶斯明天问**

2016-2017

a) Differentiate subjectivity from sentiment. How are the tasks of Subjectivity Classification and Sentiment Analysis related? [10%]

**Answer:** Subjectivity classification is often the first step for sentiment analysis

Subjective sentences do not always express positive or negative opinions

Objective sentences can express opinion indirectly

b) Give Bing Liu’s model for an opinion. Explain each of the elements in the model and exemplify them with respect to the following text, which is adapted from a TripAdvisor review of a restaurant in Sheffield. 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 and illustrate these challenges with reference to the example text. [30%]

(明天问第一句标红的句子是否主观还是客观)

“I went with my girlfriend on a Friday night, and was greeted in a friendly way by the waitress. It is simply decorated and clean, but for my personal taste was a bit too bright, and could do with a bit more colour. It is fantastic you can take your own wine and there is no uncorking fee. We was welcomed very well by the staff and I liked it that she explained the specials board to us and explained what each dish was. For starters we had the meat balls... It was amazing !! The sauce was so tasty! For our main course we had a sea food mixture with a sauce ... We felt it was a little expensive for what it was and was nice but could have been a few pounds cheaper.” Trevor M., posted 12/10/2015

**Answer:** oj is a target object

fjk is a feature of the object oj

soijkl is the sentiment value of the opinion of the opinion holder hi (usually the author of the post) on feature fjk of object oj at time tl .

oj: restaurant

fjk: environment, no fee(service), waitress, starters, main course

soijkl: negative, positive, positive, positive, negative

opinion holder hj: I, I, I, I, we

time tl: 12/10/2015

Identifying target objects

1.Named Entity Recognition: well-known tools based on gazetteers and simple context rules.

Need good gazetteers: Web is dynamic, new products appearing everyday.

Standard NE recognisers will not work for objects like names of movies

2.Bootstrap from seed gazetteers: e.g. if know that iPhone 4 is an object, can find out that iPhone 5 is also an object

Co-reference (and synonym) resolution

1.Important to resolve objects and features

c) Explain the graded lexicon-based approach for Sentiment Analysis. Given the following sentences and opinion lexicon, apply this approach to classify each sentence in S1- S3 as positive, negative or objective. Show the final emotion score for each sentence and also how this score was generated. Give any general rules that you used to calculate this score as part of your answer. Explain these rules when they are applied. [25%]

Lexicon: awesome 5

boring -3

brilliant 2

funny 3

happy 4

horrible -5

(S1) He is brilliant and funny.

(S2) I am not happy with this outcome.

(S3) I am feeling AWESOME today, despite the horrible comments from my supervisor.

**Answer:** (S1)2+3=5 positive

(S2) 4-1=3 3=>-3 negative

Negative rule: “not” is detected in neighborhood (of 5 words around);

so emotional valence of word is decreased by 1 and sign is inverted

(S3) 5+1=6 6+(-5) =1 positive

Capitalization rule: One word is capital writing.

Add +1 to positive words.

Add -1 to negative words.

1. A second approach to Sentiment Analysis is the corpus-based supervised learning approach.

(i) Explain the corpus-based supervised learning approach to Sentiment Analysis in general terms, i.e. in terms of inputs, outputs and processes involved. [5%]

(ii) 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. [10%]

(iii) Suppose you are given the following set of labelled examples as training data:

|  |  |  |
| --- | --- | --- |
| Doc | Words | Class |
| 1 | A sensitive, moving, brilliant work | Positive |
| 2 | An edgy thriller that delivers a surprising punch | Positive |
| 3 | A sensitive, insightful, beautiful film | Positive |
| 4 | Neither revelatory nor truly edgy – merely crassly flamboyant and comedically labored | Negative |
| 5 | Unlikable, uninteresting, unfunny, and completely, utterly inept | Negative |
| 6 | A sometimes incisive and sensitive portrait that is undercut by its awkward structure and . . . | Negative |
| 7 | It’s a sometimes interesting remake that doesn’t compare to the brilliant original | Negative |

Using as features just the adjectives (underlined words in the examples), how would a Naive Bayes sentiment analyser trained on these examples classify the sentiment of the new, unseen text show below?

|  |  |  |
| --- | --- | --- |
| Doc | Words | Class |
| 8 | A sensitive comedy that is moving and surprising | ??? |

Show how you derived your answer. You may assume standard pre-processing is carried out, i.e. tokenisation, lowercasing and punctuation removal. You do not need to smooth feature counts. [20%]

**Answer:**

**IR**

2013-2014

Section 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. [30%]

**Answer:** Boolean search:1. binary decision: is document relevant or not?

2. presence of term is necessary and sufficient for match

3. boolean operators are set operations (AND, OR)

Ranked algorithms: 1. frequency of document terms

2. not all search terms necessarily present in document

3. Incarnations:

• The vector space model

• The probabilistic model

• Web search engines

They should choose ranked search. Because if we want to use binary search, we should have expert knowledge then to create high-precision queries. And users should often use library. on the other hand, most users not familiar with writing Boolean queries. Users don’t want to wade through 1000s unranked result lists. There are a large number of docs. However, ranked search can start at the top and stop when satisfied

Section B:

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. [20%]

**Answer:** Stoplist removal：to exclude “non-content” words

Capitalisation: normalise all words to lower (or upper) case

Stemming: conflate morphological variants by chopping their affix

Stoplist：may at the in an , . !**(明天问，不确定)**

b) Show how Document 1, Document 2 and Document 3 would be represented using an inverted index which includes term frequency information. [10%]**(明天问这个题目是不是建立在第一小题基础上，已经进行了处理+这里的tf是不是就是我理解的还是要算比重出来)**

**Answer:**

|  |  |  |
| --- | --- | --- |
| Num | Tokens | Docs |
| 1 | sea | 1:(1,5),2:(5,9),3:(6) |
| 2 | shell | 1:(2,6),2:(6),3:(5) |
| 3 | buy | 1:(3), 2:(3) |
| 4 | my | 1:(4) |
| 5 | you | 2:(1) |
| ~~6~~ | ~~may~~ | ~~2:(2)~~ |
| 7 | lovely | 2:(4) |
| ~~8~~ | ~~at~~ | ~~2:(7)~~ |
| ~~9~~ | ~~the~~ | ~~2:(8),3:(4)~~ |
| 10 | produce | 2:(10),3:(1) |
| 11 | market | 2:(11),3:(2) |
| ~~12~~ | ~~in~~ | ~~3:(3)~~ |
| 13 | is | 3:(7) |
| ~~14~~ | ~~an~~ | ~~3:(8)~~ |
| 15 | expensive | 3:(9) |

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. [30%]

**Answer:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | sea | shell | buy | my | you | lovely | produce | market | is | expensive |
| Doc1 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Doc2 | 2 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| Doc3 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| Query1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |

Euclidean Distance:

Doc1&Query1:((2-1)2+(2-1)2+(1-0)2+(1-0)2+(0-1)2+(0-1)2)1/2=61/2

Doc2&Query1:((2-1)2+(1-1)2+(1-0)2+(1-0)2+(1-1)2+(1-1)2)1/2=31/2

Doc3&Query1:((1-1)2+(1-1)2+(1-1)2+(1-1)2+(1-0)2+(1-0)2)1/2=21/2

Doc3 is the closest document

Cosine Similarity:

Doc1&Query1:(2+2)/((4+4+1+1)1/2\*(1+1+1+1)1/2)=101/2/5=0.632

Doc2&Query1:(2+1+1+1)/((4+1+1+1+1+1)1/2\*(1+1+1+1)1/2)=5/6=0.833

Doc3&Query1:(1+1+1+1)/((1+1+1+1+1+1)1/2\*(1+1+1+1)1/2)=61/2/3=0.816

Doc2 is the closest 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. [20%]

**Answer:** Informativeness is inversely related to (document) frequency.

less common terms are more useful to finding relevant documents

more common terms are less useful to finding relevant documents

|D|/dfw: Value reduces as dfw gets larger, tending to 1 as dfw approaches |D|

Value very large for small dfw — over-weights such cases

To moderate this, take log: Inverse document frequency (idf)

idfw,D = log(|D|/dfw)

BUT Not all terms describe a document equally well

Putting it all together: tf.idf

Terms which are frequent in a document are better:

tfw,d = freqw,d

Terms that are rare in the document collection are better:

idfw,D = log (|D| / dfw )

Combine the two to give tf.idf term weighting:

tf .idfw,d,D = tfw,d · idfw,D

Most commonly used method for term weighting

The answer will be different, it will have a little influence on the appear times and it will focus on how similar about the document.

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. [20%](明天问为什么在大规模下召回率是不可能的)

**Answer:** Recall: relevant and retrieved data / relevant and retrieved data + relevant and not retrieved data = proportion of relevant documents returned

Precision: relevant and retrieved data / relevant and retrieved data + non-relevant and retrieved data = proportion of retrieved documents that are relevant

There is always a trade-off between precision and recall

F measure (also called F1):

combines precision and recall into a single figure

gives equal weight to both: F = 2PR /(P + R) P and R is the value of recall and precision

2014-2015

Section A:

a) Two Information Retrieval systems, System 1 and System 2, each return a ranked list of 10 documents they believe to be relevant for a particular query. It is known that this collection has 12 relevant documents. The following table shows whether each document returned by each system is actually relevant (X) or not (×) to the query.

Document System 1 System 2

d1 O ×

d2 × O

d3 O ×

d4 O O

d5 O O

d6 × ×

d7 × O

d8 O O

d9 × O

d10 × O

Compute the overall precision of each System, showing the equations as part of your answer. Then, compute the precision at two cutoff points: top 3 and top 5. Finally, discuss the differences between overall precision and precision at different cutoff points when comparing two or more Information Retrieval systems. Use your solution to exemplify your answer. [30%]

**Answer:**

System1: overall:5/ (5+5) =1/2

Top3:2/ (2+1) =2/3

Top5:4/ (4+1)=4/5

System2: overall:7/10

Top3:1/3

Top5:3/5

In overall precision System2 have a better performance than System1 but in the other situation System1 has a better performance. So the precision in some situations depends on the scale of the dataset. If the dataset is too small, it can’t display the right precision. Because sometimes the retrieval data will gather together and use part of the data will cause the situation like System1 is better than system2

Section B: