Homework 1

K.Frolov

Summary

For the code to work please place the original ‘data’ folder into the root folder
For code execution please run:

- ex1.ipynb - part 1
- ex2.ipynb - part 2

Part 1 - Entropy calculation

1. Entropy of sample of CVs split into words
   $H(X) = 9.79$
2. Conditional entropy of two-word phrase sample
   $H(Y|X) = 3.47$
3. Joint entropy of two-word phrase sample
   $H(X,Y) = 13.26$
4. Mutual information of 1st and second word in two-word phrases
   $I(X,Y) = 6.34$
5. Minimum possible entropy of any sample is
   $\min H(X) = 0$
   Maximum possible entropy is achieved in case of uniform distribution across the sample
   $\max H(X) = \log_2|M_X| = 11.97$
6. Magnitude of entropy characterizes “randomness” of a sample. The lower it is, the more structured the sample is.
7. Calculated entropy being less than maximum entropy indicates presence of some order in the sample.
8. Calculated entropy is greater than conditional entropy. The cause of it is presence of standard word combinations where one of the words carries the sense of the whole combination. These cases do not contribute to the total sum. Therefore, conditional entropy cannot exceed entropy.
9. Mutual information relates to the word combinations mentioned in the previous question, it consists of them exclusively. If all the words were only present in a sample accompanied by their counterpart second word would carry no information and entropy equal zero. Absence of any standard word combinations would on the contrary cause carrying the same amount of information by any word from a combination and equality of conditional entropy and entropy.

   The value of mutual information indicates that some of the words in the sample are not used without a counterpart.

   The whole sample may be encoded by 10 bits, the words which do not stand alone by 4 bits.

Part 2 - CV text sections classification

Code description

To create a base of probabilities, the provided CV corpus was split into words. Only terms containing no other symbols than letters and digits were recorded into the dictionary. Two additional features were utilized: word stemming and replacing terms containing years and months with a special term. That has proven to yield slightly better results. Further on, all the terms are recorded into the dictionary in a standard way: keys contain single words, values probabilities of those words.

The division of dataset into training and test parts was implemented in a random manner, i.e. each time the program is started exactly 80% of randomly chosen CVs are marked as training others as test. This causes the program to produce slightly different results after each restart.

For testing of the algorithm test data was organized into an array of strings which in original PDFs were separated by newline symbol. Neighbouring strings which are too short (less than 100 characters) are concatenated into one. This approach yet is not enough for complete categorization of the document - those separately classified chunks are still to be united correctly and consistently into appropriate category. The method used for categorization is Maximum Likelihood:

$$\widehat{s} = argmax_s(p(s|x)) = argmax_s(log_2 \frac{\prod_{i=1}^{N} p(s,x_i)}{\prod_{i=1}^{N} p(x_i)}) ,$$

where *N* is the length of a text chunk.

Testing results

To test the algorithm the program was run several times - each with placing of random CVs into training and test sets.

The total amount of words in the dictionary was about 17’000.

Proportions of segments:

Summary	~17%
Education	~15%
Experience	~68%

The amount of lines for classification varies from 200 to 300 from test to test. The score falls around the following values:

Segment	**Precision**	**Recall**	**F1**
Summary	0.45	0.48	0.46
Education	0.78	0.98	0.87
Experience	0.88	0.81	0.84

Conclusion

Interestingly enough, despite education’s being the average smallest segment in CVs and therefore its dictionary containing the least amount of words it is recognized quite well. This is definitely caused by presence of a significant amount of segment-specific terms such as ‘University’ etc. The structure of Summary, on the contrary, is usually less determined. It carries more entropy and often omitted, and therefore is harder to detect using the implemented method. One immediate solution is marking all text chunks which cannot be classified as one of the other two segments with sufficient confidence as summary.

Due to the reasons mentioned above the application of the algorithm to randomly structured CVs produces relatively good result. It is worth noting that the result is affected by low quality parsing of documents of different structure. Apart from increasing of the amount of training data, the quality may be improved by doing calculations based not only on single words but on n-grams as well.

Another approach to the same task may be realized by comparing the probabilities of words in a particular vocabulary not to corpus vocabulary but to other ones. In this way, the likelihood of a term belonging to a segment may be increased by unlikelihood of its belonging to another one.