FMCS1 Lab Session 4

Manuel Marques-Pita and Frank Keller

March 8, 2005

1 Introduction

During this lab session you will be interacting with a query system which is capable of retrieving properties of a big number of text documents. In this lab you will work with the plot summaries for a large number of films. You will use this system to count occurrences of different words and types of word and use this information to calculate the probabilities of certain (basic) events.

2 Task 1: Basic Probabilities

Log in your DICE account, open a terminal window and type the following line:

```
> export CORPUS_REGISTRY=/group/corpora/public/corpus_workbench/registry
```

This is telling your system where the documents you will be working with are located. Now start the Query Processor System (`cqp`) by typing:

```
> cqp -e
```

There is a small set of document groups you can access from here. In order to work with the plot summaries load the IMDB-CORPUS by typing `IMDB-CORPUS;` in the command line like this:

```
[no corpus] > IMDB-CORPUS;
```

2.1 The sample space

Our sample space for this lab consists of all the words available in the set of documents (corpus). In order to find out how many words are there in the IMDB corpus, you will need to do the following

`tokens = [];` (This will save into the variable `tokens` all the occurrences of any word)

We can then retrieve the size of this set by doing

```
size tokens;
```

2.2 Counting word types

We can now count the number of words which are *nouns* of any kind (singular, plural, etc) by doing

`nouns = [pos="N.*"];` and then retrieving the size of this set as before. `"N.*"` means any noun. If you want to count any type of verb, adjective, adverb or determiner you would use `"V.*", "JJ.*", "RB.*"` and `"DT"` respectively.

2.3 Question 1

Calculate (in a MatLab .m file) the probability that a word picked randomly from the corpus is a member of the following categories

1. noun
2. verb
3. adjective
4. adverb
5. determiner
6. none of the above

3 Task 2: Conditional Probabilities

Lets consider now the word *"flies"*. In plot summaries we could find sentences such as *"she flies through the sky like a fairy"* or *"There were so many flies in her house!"* in the first case the word is used as an active verb and in the other as a noun.

We can retrieve the instances of a word used as a certain type of word (any verb, for example) by running a query like this

`query = [word="word"%c and pos="V.*"];` (The %c means 'query is case insentive')

3.1 Question 2

If we pick the word *flies* from the corpus, what is the probability that it is used as a verb? Also, what would be the probability it is used as a noun? (add your calculations again to your MatLab file)

4 Task 3

Now, using the simplified Bayes Theorem and MatLab answer the following questions:

4.1 Question 3

What is the probability that a verb picked randomly from the corpus is the verb *flies*?

4.2 Question 4

What is the probability that a noun picked randomly from the corpus is the noun *flies*?

4.3 Question 5

What is the probability that the word *flies* is a noun or a verb?

4.4 Question 6

Devise a case similar to the *flies* case and calculate all the probabilities as in Questions (2), (3), (4) and (5). Can you think of a word which can be a noun a verb and and adverb? Interpret your answer and add your explanation to your .m file.

NOTE. You must submit your .m file with all your answers at the end of this lab session