Taalmodellen Assignment 2

Eszter Fodor, Sharon Gieske & Jeroen Rooijmans
5873320, 6167667

April 16, 2013

1. The 10 most frequent bigrams:

```
= 10 most frequent 2-grams =
('STOP', 'START') 8760
('of', 'the') 2507
('to', 'be') 2232
('in', 'the') 1917
('I', 'am') 1365
('of', 'her') 1264
('to', 'the') 1142
('it', 'was') 1010
('had', 'been') 995
('she', 'had') 978
```

2. The additional file *ngrams.txt* consists of the following 3 lines:

```
I do not
of the
She was
```

   The corresponding conditional probabilities are:

```
= conditional probabilities =
('of', 'the')
('of',)
P(the|['of']) = 0.139673519416
('She', 'was')
('She',)
P(was|['She']) = 0.184105202973

= conditional probabilities =
('I', 'do', 'not')
('I', 'do')
P(not|['I', 'do']) = 0.72972972973
```

3. The second additional file contains the two sentences:

```
Between them it was more the intimacy of sisters
Very much to the honour of both was the handsome reply
```

The corresponding probabilities are:

```
= sentence probabilities =
P(['Between', 'them', 'it', 'was', 'more', 'the', 'intimacy', 'of',
'sisters']) = 7.40518043354e-22

P(['Very', 'much', 'to', 'the', 'honour', 'of', 'both', 'was', 'the',
'handsome', 'reply']) = 1.68905697774e-22
```

4. Set A:

```
-- 2 most frequent occurance A --
P(['She', 'was', 'the', 'two', 'youngest', 'daughters', 'of',
'the']) = 2.20491955865e-12

P(['She', 'was', 'the', 'two', 'of', 'the', 'youngest',
'daughters']) = 2.92309658391e-13
```

Set B:

```
-- 2 most frequent occurance B --
P(['She', 'was', 'the', 'youngest']) = 1.30410951185e-07
P(['was', 'She', 'youngest', 'the']) = 0
```