

# Natuurlijke taalmodellen en interfaces

Eszter Fodor & Sharon Gieske

April 9, 2012

## Question 1

*grep '^A:'* Deze command zoekt naar een regel die met de empty string (^) begint gevolgd door de letter A en dubbelepunt.

*sed 's/^A: //'* Deze command verwijdert de empty string gevolgd door de letter A en een dubbelepunt (^A:) van de door grep gevonden regels. De command *j ovis-trainset.txt* definieert de output file.

## Question 2

*grep -v '^\$'* Deze command zoekt naar niet lege regels. *-v* betekent dat er juist *niet* naar de gespecificeerde regels moet worden gezocht, in dit geval een lege regel (^\$).

*sed 's/\s \+ /\n/g'* Verwijdert aan het eind van alle regels de spatie (\s) en maakt er een newline van.

*uniq -c*: Haalt duplicates uit de input en telt (*-c*) hoeveel keer iedere regel voorkomt.

*sort -g -r -k 1*: Sorteert de resultaten aan de hand van de hoeveelheid voorkomens en print deze in reversed volgorde (van het hoogste getal naar het laagste). *-g*: sorteer aan de hand van getallen, *-r*: reverse de volgorde van sorteren, *-k 1*: sorteer op de eerste key.

## Question3

Zipf distribution. ??

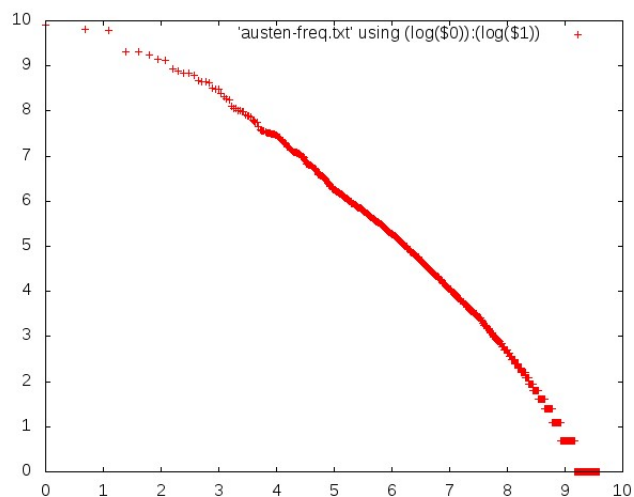
## Question 4

Met de volgende opeenvolgende commands kan de word-frequency van austen.txt geplotted worden:

```
less austen.txt | sed 's/\s\+/\n/g' | grep -v '^$'  
    | sort | uniq -c | sort -g -r -k 1  
> austen-freq.txt
```

```
gnuplot  
gnuplot> plot 'austen-freq.txt'  
          using (log($0)):(log($1))  
          gnuplot>set terminal jpeg  
gnuplot> set output 'austen.jpg'  
gnuplot> plot 'austen-freq.txt' using  
          (log($0)):(log($1))
```

Als output plaatje krijgen we dan:



De twee plaatjes van de word-frequencies lijken vrijwel identiek.

## Question 5

### 10 most frequent bigrams with their frequen:

1487, ik wil  
514, dank u  
399, nee dank  
390, wil van  
279, den haag  
264, tien uur  
249, u wel  
223, acht uur  
212, wil om  
205, negen uur

### 10 most frequent trigrams with their frequencies:

393, nee dank u  
387, ik wil van  
248, dank u wel  
211, ik wil om  
164, nee ik wil  
153, ik wil graag  
141, om tien uur  
128, van den haag  
117, ja dat klopt  
117, om acht uur  
117, ja dat klopt  
117, om acht uur

### 10 most frequent tetragrams with their frequencies:

177, nee dank u wel  
81, ik wil reizen van  
63, ik wil graag van  
53, van den haag centraal  
50, den haag centraal naar  
46, nee hoor dank u  
46, nee dat hoeft niet  
46, nee hoor dank u  
46, nee dat hoeft niet  
45, van den haag naar  
42, dat is niet nodig  
41, ik wil van den