Answers Assignment A (Step 2)

Name: Alex Khawalid, 10634207
Philip Bouman, 10668667

Introductie

In deze opdracht zullen we proberen de kansen te berekenen op het voorkomen van een woord of zin. Deze analyse is gebasseerd op n-gram en
(n-1)-gram tables geconstrueerd uit het corpus. Daarnaast zullen we ook kijken naar permutaties van verschillende sets en de kansen van voorkomen van deze permutaties in het corpus.

Probleem beschrijving

Als eerste zullen we uit het corpus de n-gram en (n-1)-gram tables moeten creëren. Daarnaast zal het programma een aantal extra input files moeten kunnen lezen en deze eventueel kunnen labelen. Als laatste zal het programma de kansen moeten kunnen bereken op het voorkomen van woorden of woordsets en de permutaties hiervan.

Aanpak

Om het begin en einde van de zinnen af te bakenen, wordt er aan het begin van elke zin 'START' en aan het einde van elke zin 'END' toegevoegd. Vervolgens construeert het programma de n-gram en (n-1)-gram tables uit het corpus (austen.txt). In de tweede stap worden uit een text-file zinnen uitgelezen en de kans bepaald dat deze in het corpus voorkomen. In de derde stap van het programma worden de kansen van de voorgaande woorden gecombineerd gebasseerd op de eerder gecreërde n-gram tables. In de laatste stap van het programma worden van twee sets (a & b) alle permutaties gegenereerd en wordt er vervolgens gekeken wat de kans is dat deze permutaties in het corpus voorkomen.

Resultaten

1. Ten most frequent bigrams in austen.txt:

	$n = 2$	propability
1	END START	61322
2	START and	3917
3	of the	2433
4	to be	2194
5	START to	2051
6	START of	2022
7	START I	1997
8	in the	1872
9	START the	1380
10	I am	1338

4. Twee permutaties met de hoogste kans:

   Set A = {know,I,opinion,do,be,your,not,may,what}:
   1. I do not know what your opinion may be
   2. your what opinion not may know do be I

   Set B = {I,do,not,know}:
   1. I do not know
   2. not know do I