CS4442 – Assignment 2

Dustin Dobransky 250575030

1. This question required us to compare n-grams between two texts.
   1. For Dostoyevsky’s works, the results were as follows:

|  |  |
| --- | --- |
| nGram size | Percentage of works in 2 not in 1 |
| 1 | 40.647 |
| 2 | 72.681 |
| 3 | 89.796 |
| 4 | 97.606 |
| 5 | 99.536 |
| 6 | 99.891 |

The largest common n-gram occurs at n = 17, with “repulsion that s what i m afraid of that s what may be too much for me”

* 1. For Dickens vs. Kafka’s works, the results were as follows:

|  |  |
| --- | --- |
| nGram size | Percentage of works in 2 not in 1 |
| 1 | 32.880 |
| 2 | 77.458 |
| 3 | 95.433 |
| 4 | 99.195 |
| 5 | 99.885 |
| 6 | 99.977 |

The largest common n-gram occurs at n = 7 with two n-grams:

1. “in the middle of the table and”
2. “there is no such thing as a”
   1. For “MarxEngelsManifest” vs. “SmithWealthNations”, the results were:

|  |  |
| --- | --- |
| nGram size | Percentage of works in 2 not in 1 |
| 1 | 84.185 |
| 2 | 97.468 |
| 3 | 99.534 |
| 4 | 99.920 |
| 5 | 99.987 |
| 6 | 99.998 |

The largest common n-grams occur at n = 6 with:

1. of nature and of reason the
2. is the same as that of
3. to keep up the rate of
4. in order to keep up the
5. of a man s own labour
6. from them what they have not
   1. Discussion:

From the results in a), it can be seen that books written from the same author will have a large number of similar n-grams. This is expected, as the author has a distinct writing style, which will be reflected in their works. In this case, similarity in writing style is identified by the recurrence of n-grams across different novels.

For the remaining two tests, both were authored by different people, and the differentiation between n-grams was large. Past 6-grams, the works did not contain significant overlap.