

6.1 Compute Levenshtein Distance:

② Sasquatchuan → Saskatchewan = 2

↓ a ↓ e
Replace e → a
Replace u → e

③ Jacquiri → Jaiquiri =

Replace c → i

④ hors d'œuvres → hors d'œuvres = 4

↑
insert ' → d'œuvres
↑
insert o → d'œuvres
↑
Replace r to u → d'œuvres
↑
insert r → d'œuvres

6.2

1. I am Sam. I am Sam. Sam-I-Am.
2. That Sam-I-Am! That Sam-I-Am! I do not like that Sam-I-Am!

① TF-IDF for word "Sam" in Doc1

$$TF = \begin{cases} 1 + \log(f_{t,d}) & \text{if } f_{t,d} > 0 \\ 0 & \text{other} \end{cases} \times \log\left(\frac{N}{N_t}\right)$$

$$= (1 + \log(3)) \times \log\left(\frac{10}{3}\right) = 1.928$$

② "cats" in Doc1

$$(1 + \log(0)) \times \log\left(\frac{10}{3}\right) = 0$$

③ "like" in Doc10

$$(1 + \log(1)) \times \log\left(\frac{10}{8}\right) = 0.223$$

Exercises

3. Do you like green eggs and ham?
4. I do not like them, Sam-I-Am.
5. I do not like green eggs and ham.
6. Would you like them here or there?
7. I would not like them here or there.
8. I would not like them anywhere.
9. I do not like green eggs and ham.
10. I do not like them, Sam-I-Am.

6.3

6.3 Consider three answers of smokers who explained how they were able to quit smoking:

- "Novels. Particularly the book author Allan Carr."
- "A Book! A good book. A novel idea"
- "The Easy Way by Allan Carr. (Self Help books)"

① novel particularly book author allan Carr 6

book good book novel idea 5

easy way allan Carr self help book 7

Create n-gram matrix from here:

	Doc 1	Doc 2	Doc 3
Novel	1	1	0
Particularli	1	0	0
book	1	2	1
author	1	0	0
allan	1	0	1
Car	1	0	1
good	0	1	0
idea	0	1	0
easy	0	0	1
way	0	0	1
self	0	0	1
help	0	0	1
# words	6	5	7

⑥ List all bigrams that can be created from the text that occur in more than one answer

allan - car

⑦ Compute TF-IDF for "book" for sentence "a book! probably a good book."

stemming → ~~a~~ book probabl ~~a~~ good book

$$TF = 1 + \log(f_{t,d}) = 1 + \log(2)$$

$$IDF = \log\left(\frac{N}{N_t}\right) = \log\left(\frac{3}{3}\right) = \log(1) = 0$$

$$TF-IDF = 0$$