AI6122 Text Data Management and Processing

Qixuan Bi
biqi0002@e.ntu.edu.sg
Nanyang Technological University
Singapore

Ruoxi Fan
rfan002@e.ntu.edu.sg
Nanyang Technological University
Singapore

Zhang Wan
zwan004@e.ntu.edu.sg
Nanyang Technological University
Singapore

Rui Wang
wang1806@e.ntu.edu.sg
Nanyang Technological University
Singapore

Junyu Yin
s210027@e.ntu.edu.sg
Nanyang Technological University
Singapore

ABSTRACT

1 INTRODUCTION

2 DATASET ANALYSIS

In this project, we use *Amazon product data* from [1] to conduct all the following experiments. Specifically, we randomly select 200 product reviews from each of the two 5-core datasets with categories "**Health and Personal Care**" and "**Video Games**", respectively.

2.1 Writing Sytle

2.2 POS Tagging

2.3 Sentence Segmentation

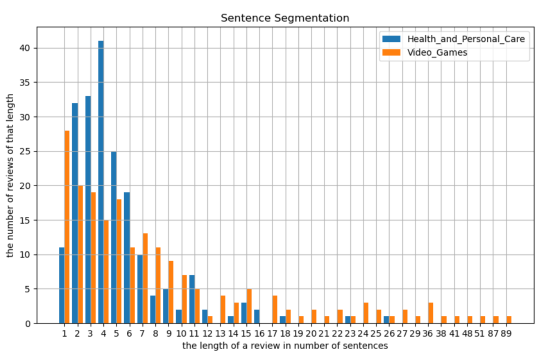


Figure 1: The distribution of the two datasets in terms of the number of sentences.

2.4 Tokenization and Stemming

2.5 Indicative Words

3 SEARCH ENGINE

4 REVIEW SUMMARIZER

5 SENTIMENT ANALYSIS

CONTRIBUTIONS

Junyu Yin: Dataset Analysis.

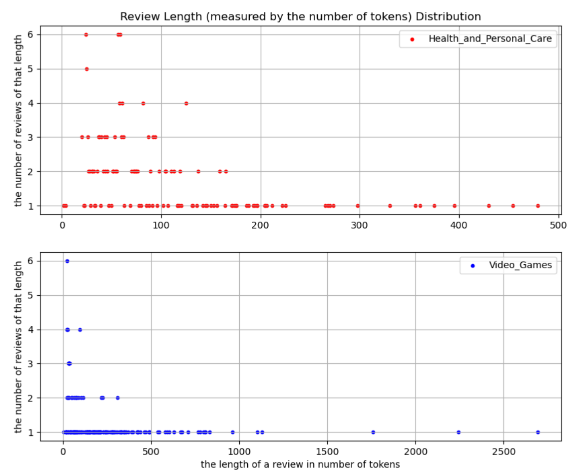


Figure 2: The distribution of the two datasets in terms of the number of tokens.

REFERENCES

[1] Ruining He and Julian McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *WWW*.

It	*helped*	*me*	*take*	*off*	*those*	*extra*	*pounds*	*that*	*I*	*had*	*gained*	*during*	*the*
PRP	VBD	PRP	VB	RP	DT	JJ	NNS	IN	PRP	VBD	VBN	IN	DT
holidays	.												
NNS	.												
Everyone	*should*	*read*	*this*	*article*	*and*	*use*	*this*	*formula*	.				
NN	MD	VB	DT	NN	CC	VB	DT	NN	.				
Pretty	*much*	*the*	*only*	*way*	*to*	*shave*	*cheaper*	*is*	*to*	*use*	*a*	*double*	*edge*
RB	RB	DT	JJ	NN	TO	VB	JJR	VBZ	TO	VB	DT	JJ	NN
razor	,	*and*	*even*	*then*	*you*	*do*	*n't*	*save*	*that*	*much*	*more*	.	
NN	,	CC	RB	RB	PRP	VBP	RB	VB	RB	JJ	RBR	.	
Right	*now*	*the*	*only*	*things*	*I*	*'ve*	*unlocked*	*are*	*Backpack*	*option*	,	*Koga*	,
RB	RB	DT	JJ	NNS	PRP	VBP	VBN	VBP	NNP	NN	,	NNP	,
Kikyo	,	*Kagura*	,	*and*	*you*	*already*	*have*	*Inuyasha*	,	*Kagome*	,	*Miroku*	,
NNP	,	NNP	,	CC	PRP	RB	VBP	NNP	,	NNP	,	NNP	,
Shippou	,	*and*	*Sango*	*when*	*the*	*game*	*starts*	*off*	.				
NNP	,	CC	NNP	WRB	DT	NN	VBZ	RP	.				
My	*first*	*xbox*	*since*	*2004*	.								
PRP$	JJ	NN	IN	CD	.								

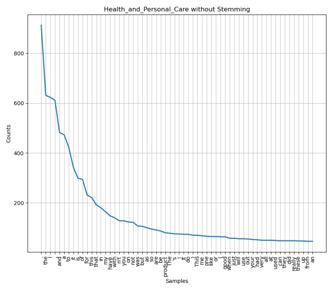
Table 1: The POS tagging results.

	w/o stemming	Porter stemming	Lancaster stemming	Snowball stemming
Health and Personal Care	3546	2591	2372	2559
Video Games	6582	4490	4049	4409

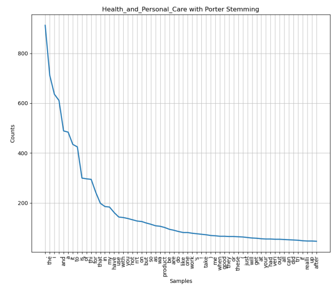
Table 2: The number of unqiue tokens with and without stemming.

Health and Personal Care	product	skin	smell	products	used	brand	bottle	taste	price	weight
Video Games	game	games	play	fun	graphics	playing	played	story	system	gameplay

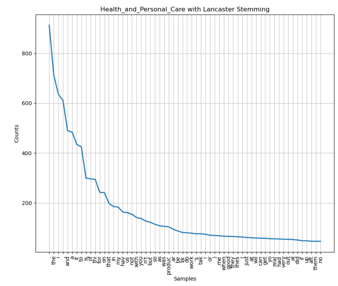
Table 3: The top-10 most indicative words in each of the two datasets.



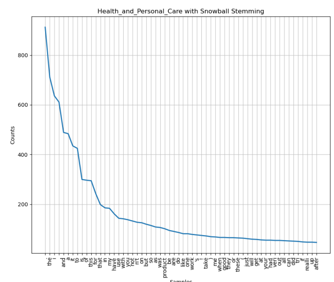
(a) Without stemming



(b) Porter stemming

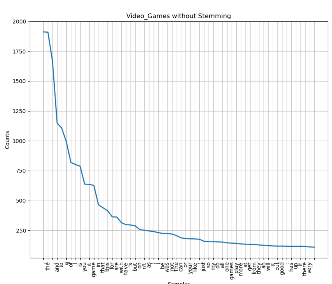


(c) Lancaster stemming

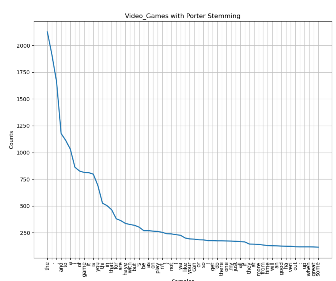


(d) Snowball stemming

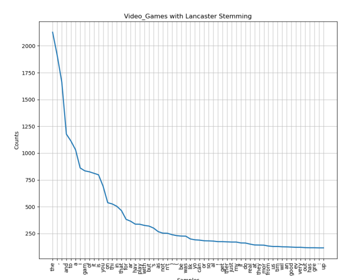
Figure 3: Word frequency distribution of *Health and Personal Care* with and without stemming.



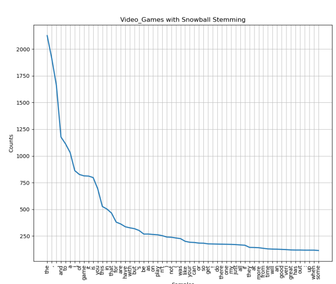
(a) Without stemming



(b) Porter stemming



(c) Lancaster stemming



(d) Snowball stemming

Figure 4: Word frequency distribution of *Video Games* with and without stemming.