Text Analytics: Practical 2

Question 1 Statement for analysis

A Beautiful Mind is biographical film based on life of the American mathematician John Nash, an Economics Nobel and Abel winner. He won J. V. Neumann Theory Prize for discovering non-cooperative equilibria. Reference: https://en.wikipedia.org/wiki/John_Forbes_Nash_Jr. In 1951, Massachusetts Institute of Technology (MIT) hired Nash as C. L. E. Moore instructor in mathematics faculty.

1(a) Output of nltk.word_tokenize():

A	Beautiful	Mind	is	biographical
film	based	on	life	of
the	American	mathematician	John	Nash
,	an	Economics	Nobel	And
Abel	winner	.	He	won
J.	V.	Neumann	Theory	Prize
for	discovering	non-cooperative	equilibria	.
Reference	:	https	:	//en.wikipedia.org/ wiki/John_Forbes _Nash_Jr
.	In	1951		Massachusetts
Institute	of	Technology	(	i.e
.	MIT	)	hired	Nash
as	C.	L.	E.	Moore
instructor	in	mathematics	faculty	.

Oddities

Following classes of oddities are noticed

- URLs are not successfully noticed as one unit of speech. Thus, URL in text was split into 3 components “https”, “:” and “//en.wikipedia.org/wiki/john_forbes_nash_jr”. To counter this challenge, regex can be used to define a URL format.
- Abbreviation like “i.e.” are not correctly handled. To tackle this situation, punctuation tokenizer can be used which accepts library of abbreviations and handle occurrences.

1(b) 1(c) Output of Tagging Parts of Speech (POS Tagging)

text_token	PoS_tag	text_token	PoS_tag	text_token	PoS_tag
a	DT	the	DT	winner	NN
beautiful	JJ	american	JJ	.	.
mind	NN	mathematician	NN	he	PRP
is	VBZ	john	NN	won	VBD
biographical	JJ	nash	NN	j.	NN
film	NN	an	DT	v.	NN
based	VBN	economics	NNS	neumann	JJ
on	IN	nobel	NN	theory	NN
life	NN	and	CC	prize	NN
of	IN	abel	NN	for	IN

Text Analytics: Practical 2

text_token	PoS_tag	text_token	PoS_tag	text_token	PoS_tag
discovering	VBG	1951	CD	nash	JJ
non-cooperative	JJ	massachusetts	FW	as	IN
equilibria	NN	institute	NN	c.	JJ
.	.	of	IN	l.	NN
reference	NN	technology	NN	e.	NN
:	:	(	(	moore	NN
https	NN	i.e	JJ	instructor	NN
:	:	.	.	in	IN
//en.wikipedia.org /wiki/john_forbes _nash_jr	NN	mit	NN	mathematics	NNS
.	.	)	)	faculty	NN
in	IN	hired	VBD	.	.

Abbreviations for Parts of Speech identified by POS Tagger:

CC: coordinating conjunction	CD: cardinal digit	EX: existential there (e.g: "there is")
FW: Foreign word 'Massachusetts'	IN: Preposition/ subordinating conjunction	JJ: adjective 'huge
JJR: adjective comparative	JJS: adjective, superlative 'best	LS: list marker
MD: modal could, will	NN: noun, singular 'boy'	NNS: noun plural 'birds'
POS: possessive ending	PRP: personal pronoun 'she, her'	PRP$: possessive pronoun
NNP: proper noun, singular	NNPS: proper noun, plural	PDT: predeterminer 'all the kids'
RP: particle give up	TO: to go 'to' the store.	UH: interjection, ohhh
RB: adverb very, silently	RBR: adverb, comparative better	RBS: adverb, superlative best
VB:verb, base form take	VBD: verb, past tense took	VBG: verb, gerund/present
VBN: verb, past participle	VBP : verb, sing. present, non-3d take	VBZ: verb, 3rd person sing. present
WDT: wh-determiner which	WP: wh-pronoun who, what	DT: Determiner

Table Reference : www.sketchengine.eu/tagsets/penn-treebank-tagset/

Oddities:

- (John,NN) (nash,NN) should be classified as Proper Singular Noun (NNP)
- ('ie', 'JJ'), is classified as adjective instead of abbreviation
- ('c.' , JJ) should be classified as a noun(NN), just like ('l.' , NN), ('e.' , NN)
- (nash, JJ) should be classified as Proper Singular Noun (NNP). This is double oddity since "nash"has also been tagged as (nash,NN) earlier
- ('mathematics', NNS) word mathematics is tagged as plural noun, though it is singular noun (NN) in context of sentence.

Text Analytics: Practical 2

Question 2 Statement for analysis

After financial crisis of 2007–2008, the Spanish economy plunged into recession, entering negative macroeconomic performance cycle. Compared to the EU's average, the Spanish economy entered recession later (the economy was still growing by 2008), but it stayed there longer. The economic boom of the 2000s was reversed.

1(a) Output of Porter Stemming

text_token	porter_stemming	text_token	porter_stemming	text_token	porter_stemming
After	after	the	the	it	it
financial	financi	EU	EU	stayed	stay
crisis	crisi	's	's	there	there
Of	of	average	averag	longer	longer
2007-2008	2007-2008	the	the	.	.
The	the	Spanish	spanish	The	the
Spanish	spanish	economy	economi	economic	econom
economy	economi	entered	enter	boom	boom
plunged	plung	recession	recess	of	of
Into	into	later	later	the	the
recession	recess	(	(	2000s	2000
entering	enter	the	the	was	wa
negative	neg	Economy	economi	reversed	revers
macroeconomic	macroeconom	was	wa	.	.
performance	perform	still	still		
cycle	cycl	growing	grow		
.	.	by	by		
Compared	compar	2008	2008		
to	to	but	but		

Observations for Porter Stemmer:

Porter stemmer was designed to prune morphological or inflection endings [-ed, -ing, -es etc.] of words. So essentially it removes suffix from word. Purpose of stemming is to bring variant forms of a word [1]; but often words not belonging English vocabulary are generated which add up to noise in data. Such anomalies are listed below:

- Financial → financi
- Crisis → crisi
- Economy → economi
- negative →neg
- cycle → cycl
- was → wa
- reversed → revers

Reference:

[1] https://tartarus.org/martin/PorterStemmer/

2(b) Output of WordNet Lemmatizer:

text_token	PoS_tag	lemmantized_word	text_token	PoS_tag	lemmantized_word
after	IN	after	entered	VBD	enter
financial	JJ	financial	recession	NN	recession
crisis	NN	crisis	later	RB	later
of	IN	of	(	(	(
2007-2008	CD	2007-2008	the	DT	the
the	DT	the	economy	NN	economy
spanish	JJ	spanish	was	VBD	be
economy	NN	economy	still	RB	still
plunged	VBD	plunge	growing	VBG	grow
into	IN	into	by	IN	by
recession	NN	recession	2008	CD	2008
entering	VBG	enter	)	)	)
negative	JJ	negative	but	CC	but
macroeconomic	JJ	macroeconomic	it	PRP	it
performance	NN	performance	stayed	VBD	stayed
cycle	NN	cycle	there	RB	there
.	.	.	Longer	RBR	longer
compared	VBN	compare	.	.	.
to	TO	to	the	DT	the
the	DT	the	economic	JJ	economic
eu	NN	eu	boom	NN	boom
's	POS	's	of	IN	of
average	NN	average	the	DT	the
,	,	,	2000s	CD	2000s
the	DT	the	was	VBD	be
spanish	JJ	spanish	reversed	VBN	reverse
economy	NN	economy	.	.	.

WordNet Lemmatizer has improved performance against Porter Stemming. This happens since it considers part of speech tag before pruning suffixes. POS tag helps context to bring out context of word.

The WordNet Lemmatization still fails in some instances such as:

- Word ("longer", RBR) is not reduced to root "long".
- Word ("Stayed", VBD) is not reduced to root verb "Stay".

Text Analytics: Practical 2

2(c) Comparison between Porter Stemmer and WordNet Lemmatizer

text_token	porter_stemming	lemmantized_word
financial	financi	financial
crisis	crisi	crisis
economy	economi	economy
plunged	plung	plunge
recession	recess	recession
entering	enter	enter
negative	neg	negative
macroeconomic	macroeconom	macroeconomic
performance	perform	performance
cycle	cycl	cycle
Compared	compar	compare
EU	EU	eu
average	averag	average
economy	economi	economy
entered	enter	enter
recession	recess	recession
economy	economi	economy
was	wa	be
growing	grow	grow
stayed	stay	stayed
economic	econom	economic
was	wa	be
reversed	revers	reverse

Both Lemmatizer and Stemmer algorithms reduce words to root form. However, Stemming is a heuristic process that chops off the ends of words and often includes the removal of derivational affixes. Whereas, Lemmatization uses a vocabulary [dictionary of root words] and morphological analysis of words [Analyze Parts of Speech]. It removes inflectional endings only and returns the base or dictionary form of a word, termed "lemma" [1]. So, Lemmatizer is a **better** choice between the two. Due to added dictionary lookup for lemma [canonical form/ root word], false negative rate is improved in Lemmatizer as against Stemming.

Also, stemming commonly collapses derivationally related words, whereas lemmatization commonly only collapses the different inflectional forms of a lemma. For example, in text above :

- Stemming(Recession) → recess *[reduced to lemma]*
- Lemmatize(Recession) → recession *[derivation preserved]*

Reference:

[1] nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html