TF-IDF

Ko, Youngjoong

Sungkyunkwan University

nlp.skku.edu, nlplab.skku.edu

成均館大學校
SUNG KYUN KWAN UNIVERSITY
1398

Contents

❖ NLTK Part-Of-Speech(POS) Tag

❖ Count based Word Representation

- Bag of Words(BoW)
- TF-IDF

❖ Assignment

POS Tag

❖ POS Tag

- A POS tag is a label assigned to each word in a text to indicate the part of speech.
  Ex) subject, verb, object, etc.

- In general, main components of sentences are subject, verb, object, and complement, and these are usually verbs or nouns.

- In this assignment, we use only verbs and nouns tags.
  - Verb : (VB, VBD, VBG, VBN, VBP, VBZ)
  - Noun : (NN, NNS, NNP, NNPS)

  (Details in page 4)

POS Tag

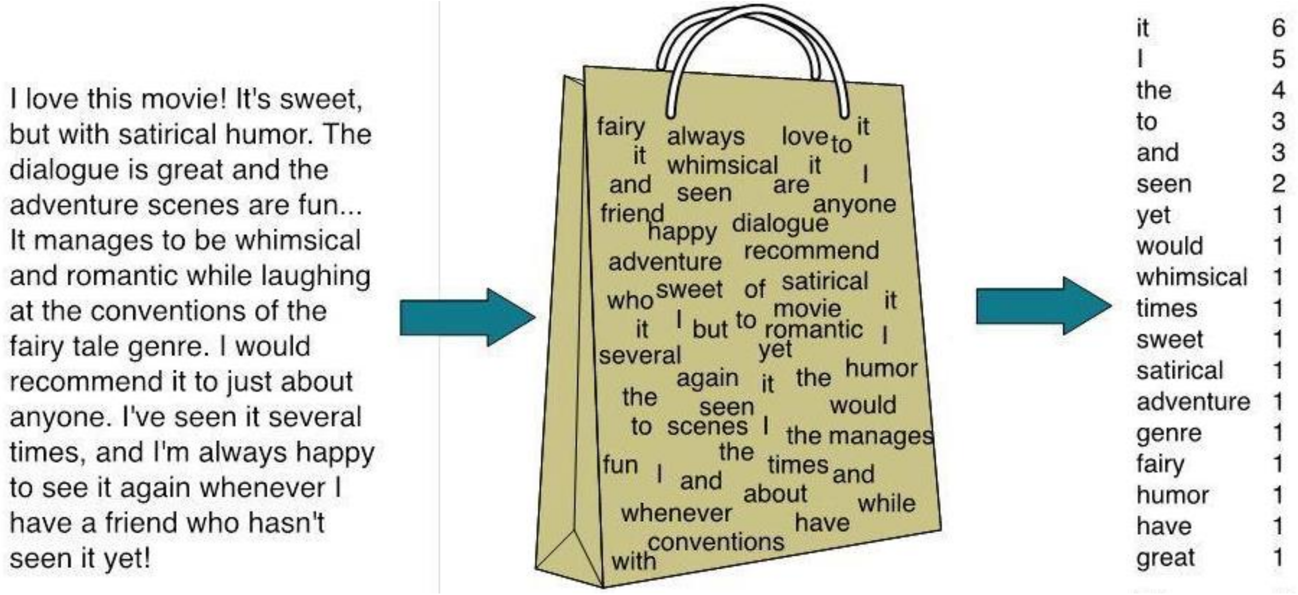
❖ NLTK POS Tag

Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	*there* is
FW	foreign word	d’hoevre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	big
JJR	adjective, comparative	bigger
JJS	adjective, superlative	biggest
LS	list marker	1)
MD	Modal	could, will
NN	noun, singular or mass	Door
NNS	noun plural	Doors
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	Predeterminer	*both* the boys
POS	possessive ending	friend*‘s*
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his
RB	Adverb	however, usually
RBR	adverb, comparative	Better
RBS	adverb, superlative	Best
RP	Particle	give *up*
TO	To	*to* go, *to* him
UH	Interjection	Uhhuhhuhh
VB	verb, base form	Take
VBD	verb, past tense	Took
VBG	verb, gerund/present participle	Taking
VBN	verb, past participle	Taken
VBP	verb, sing. present, non-3d	Take
VBZ	verb, 3rd person sing. Present	Takes
WDT	wh-determiner	Which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	Whose
WRB	wh-abverb	where, when

Count based Word Representation

❖ Bag of Words(BoW)

- A method of numerical expression of text data that focuses only on the frequency of words without considering the order of words.



I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1

Count based Word Representation

❖ Bag of Words(BoW) based Document Features Extraction Method

- Doc0: I/PRP am/VBP a/DT boy/NN
- Doc1: I/PRP am/VBP a/DT girl/NN

1) Remove duplicates from all words in Doc0 and Doc1, and extract verbs and nouns. List each word in column form. (Sorted by alphabet)

   - [‘am/VBP‘, ‘boy/NN’, ’girl/NN’]

2) Give each of the following words a unique number (index).

   - { ‘am/VBP‘ : 0, ‘boy/NN’ : 1, ‘girl/NN’ : 2 }

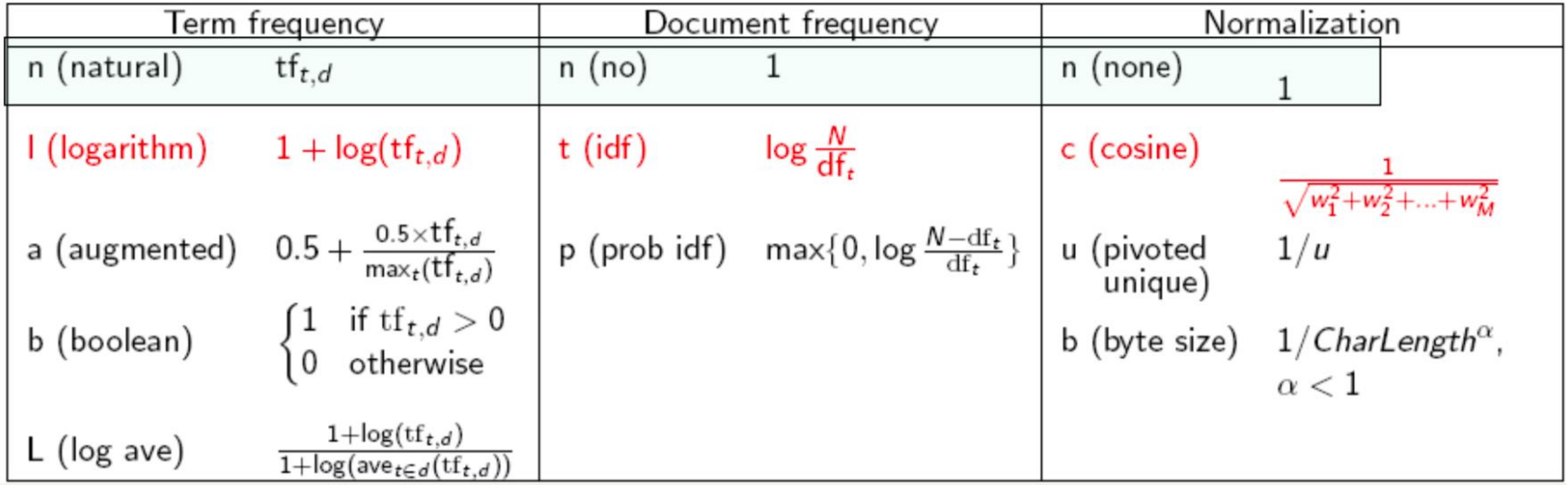
3) In each document, write the frequency in which the word appears in each number.

   - Index : 0 1 2
   - Doc0 : [1, 1, 0]
   - Doc1 : [1, 0, 1]

Count based Word Representation

❖ TF-IDF(Term Frequency-Inverse Document Frequency)

- TF(Term Frequency)
  : The number of times that term t occurs in document d ($tf_{t,d}$)
- DF(Document Frequency)
  : The number of documents that contain the term t (df_t)
- IDF(Inverse Document Frequency)
  : Inverse value of DF.



Term frequency		Document frequency		Normalization	
n (natural)	$\mathrm{tf}_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1+\log(\mathrm{tf}_{t,d})$	t (idf)	$\log\frac{N}{\mathrm{df}_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2+w_2^2+\ldots+w_M^2}}$
a (augmented)	$0.5+\frac{0.5\times\mathrm{tf}_{t,d}}{\max_t(\mathrm{tf}_{t,d})}$	p (prob idf)	$\max\{0,\log\frac{N-\mathrm{df}_t}{\mathrm{df}_t}\}$	u (pivoted unique)	$1/u$
b (boolean)	$\begin{cases}1 & \text{if } \mathrm{tf}_{t,d}>0\\ 0 & \text{otherwise}\end{cases}$			b (byte size)	$1/CharLength^{\alpha}$, $\alpha<1$
L (log ave)	$\frac{1+\log(\mathrm{tf}_{t,d})}{1+\log(\mathrm{ave}_{t\in d}(\mathrm{tf}_{t,d}))}$				

Count based Word Representation

❖ TF (Term Frequency)

- Doc0 : I/PRP am/VBP a/DT boy/NN
- Doc1 : I/PRP am/VBP a/DT girl/NN
- Doc2 : who/WP is/VBZ a/DT boy/NN

- $tf_{t,d}$: The number of times that term t occurs in document d ($tf_{t,d}$)

	am/VBP	is/VBZ	boy/NN	girl/NN
Doc0	1	0	1	0
Doc1	1	0	0	1
Doc2	0	1	1	0

Count based Word Representation

❖ IDF (Inverse Document Frequency)

- Doc0 : I/PRP am/VBP a/DT boy/NN
- Doc1 : I/PRP am/VBP a/DT girl/NN
- Doc2 : who/WP is/VBZ a/DT boy/NN
  - $log\frac{N}{df_t}$: Inverse value of DF.
  - N : Total number of document
  - df_t : The number of documents that contain the term t (df_t)

  Ex) am/VBP = $log\frac{3}{2} = 0.176$, girl/NN= $log\frac{3}{1} = 0.477$

	am/VBP	is/VBZ	boy/NN	girl/NN
IDF	0.176	0.477	0.176	0.477

Count based Word Representation

❖ TF-IDF(Term Frequency Inverse Document Frequency)

- $tf_{t,d} \times log\frac{N}{df_t}$

Ex) am/VBP : $(1 \times 0.176) = 0.176$

	am/VBP	is/VBZ	boy/NN	girl/NN
Doc0	0.176	0	0.176	0.477
Doc1	0.176	0	0.176	0
Doc2	0	0.477	0	0

Assignment

❖ Assignment

1) The attached JSON file in assignment 2 consists of 300 articles with multiple sentences.

   - Please refer to page 13 for JSON input file format.

2) The output is nouns and verbs with TF(Term Frequency) values after morpheme analysis on the given articles. (Sorted by alphabet)

   - Verbs (VB, VBD, VBG, VBN, VBP, VBZ) and nouns (NN, NNS, NNP, NNPS)

3) You MUST create an output text file.

   - Please refer to page 14 for the format of the result.

4) You can use external libraries, but you MUST use NLTK library for *tokenizer* and *POS tagger*.

Assignment

- ❖ Submit File
  1) Python code file (.py) (python version 3.x)
     - Format: “Student Number_Name_TF.py”.

       - Ex) “2020000000_홍길동.py”

  2) TEXT file (.txt)
     - Format: “Student Number_Name_TF.txt”.

JSON Input File Format

Article Topic

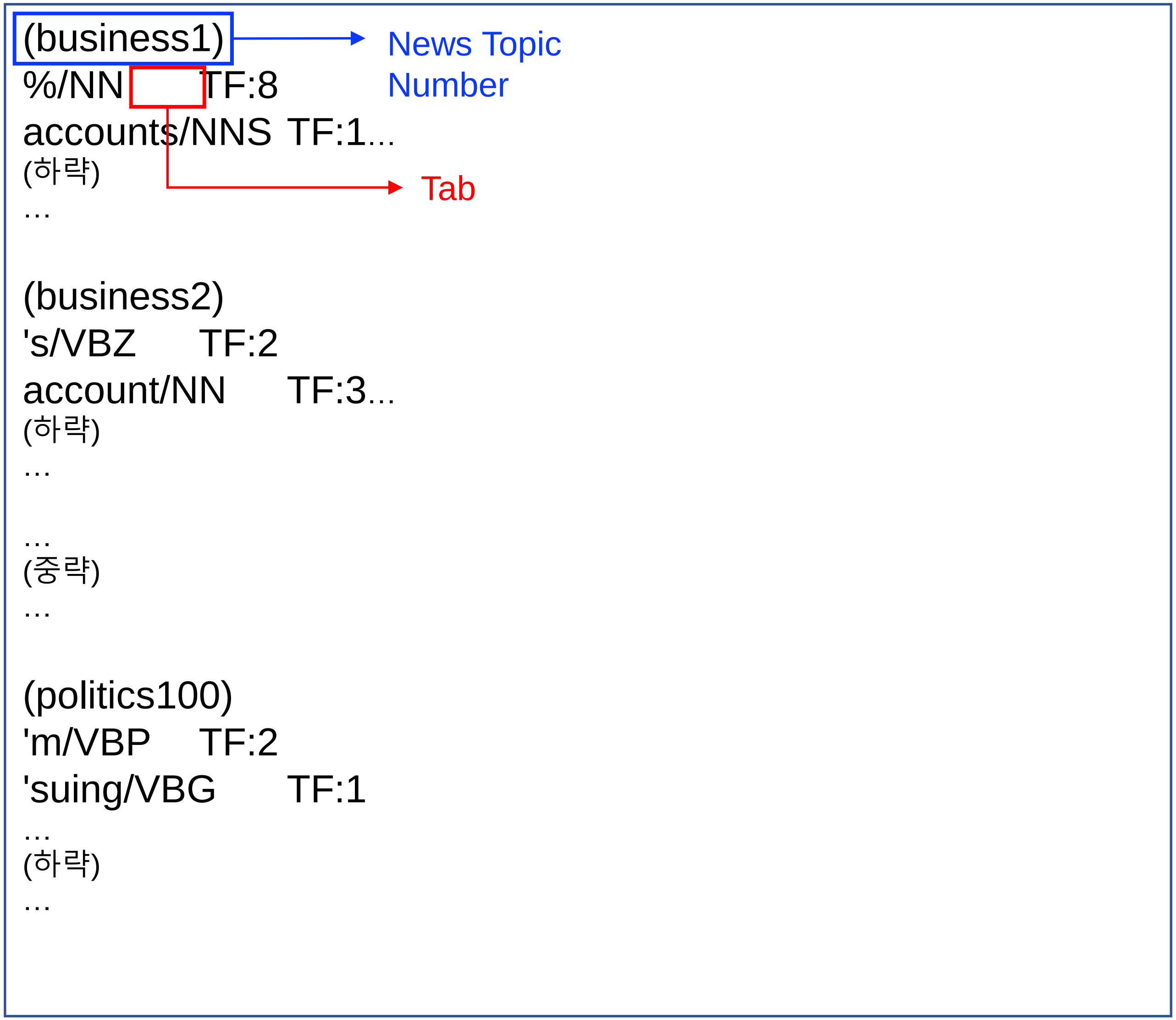
article

```
"business": [
"Ad sales boost Time Warner profit  Quarterly profits at US media …
"Dollar gains on Greenspan speech  The dollar has hit …
…
(중략)
…
"GM, Ford cut output as sales fall …
"Ebbers denies WorldCom fraud …
  ],
…
(중략)
…
”politics": [
"Labour plans maternity pay rise…
"Watchdog probes e-mail deletions…
…
(중략)
…
"Blair stresses prosperity goals…
"Guantanamo man 'suing government'…
  ]
```

- JSON input file includes 300 articles on business, politics, and tech topics.

Output File Format

Student Number_Name_TF.txt



```
(business1)
%/NN	TF:8
accounts/NNS	TF:1...
(하략)
...

(business2)
's/VBZ	TF:2
account/NN	TF:3...
(하략)
...

...
(중략)
...

(politics100)
'm/VBP	TF:2
'suing/VBG	TF:1
...
(하략)
...
```