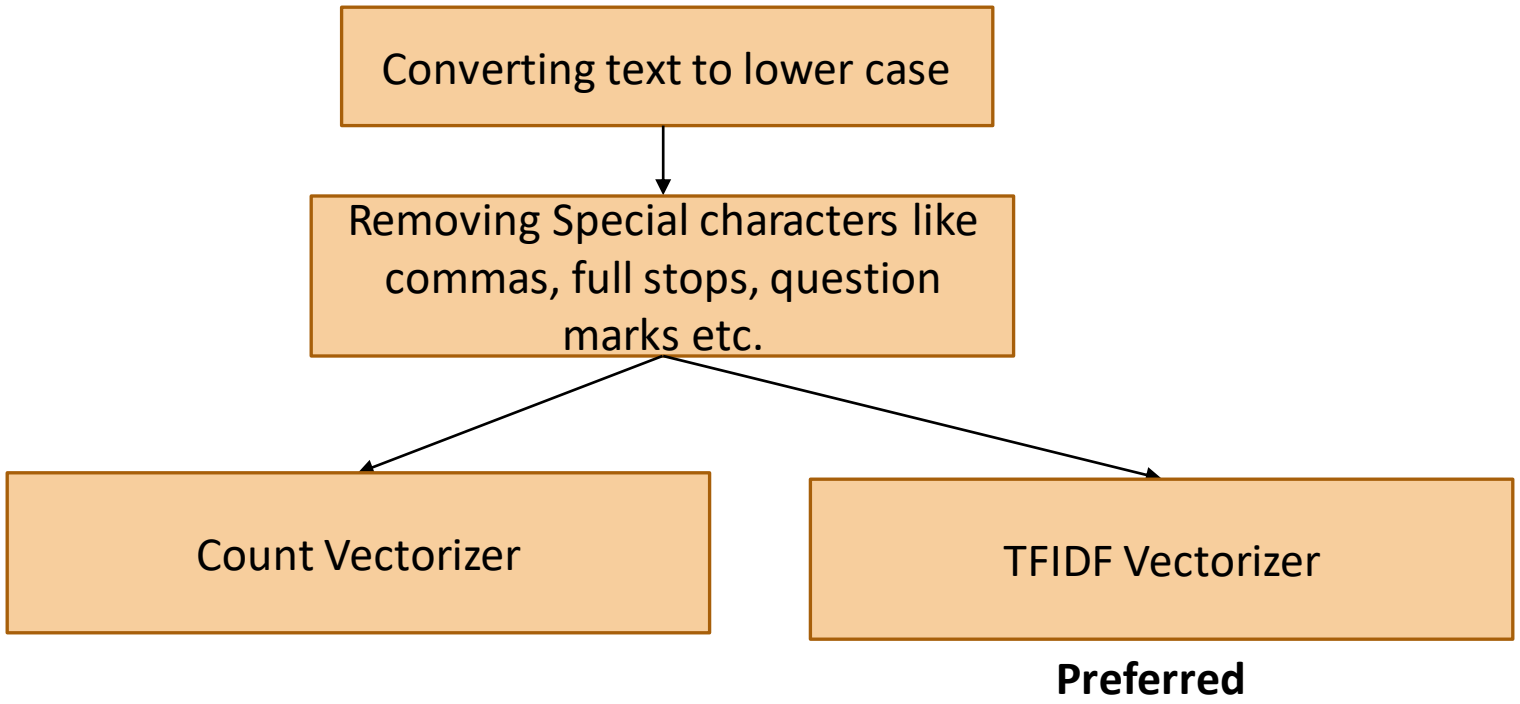
Text Classification

UTKARSH GAIKWAD

CLASS STARTING SHARP AT 10:15 AM

Text Pre-processing basic steps



Count Vectorizer example

- I love to eat pizza, I love pizza.
- Pizza is my favorite food.
- I eat pizza every Friday.

	I	love	to	eat	pizza	is	my	favourite	food	every	Friday
Sentence 1	2	2	1	1	2	0	0	0	0	0	0
Sentence 2	0	0	0	0	1	1	1	1	1	0	0
Sentence 3	1	0	0	1	1	0	0	0	0	1	1

TFIDF Vectorizer Example

- Document 1: "The quick brown fox jumps over the lazy dog"
- Document 2: "The quick brown fox is very clever and quick"
- Document 3: "The brown dog is quick and the brown fox is clever"

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

TF- Term Frequency

Document 1: "The quick brown fox jumps over the lazy dog"
Document 2: "The quick brown fox is very clever and quick"
Document 3: "The brown dog is quick and the brown fox is clever"

Term	Document 1	Document 2	Document 3
the	2	1	2
quick	1	2	2
brown	1	1	2
fox	1	1	1
jumps	1	0	0
over	1	0	0
lazy	1	0	0
dog	1	0	1
is	0	1	2
very	0	1	0
clever	0	1	1

IDF (Inverse Document Frequency)

Document 1: "The quick brown fox jumps over the lazy dog"
Document 2: "The quick brown fox is very clever and quick"
Document 3: "The brown dog is and the brown fox is clever"

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Term	IDF
the	0.0000
quick	0.1761
brown	0.1761
fox	0.4055
jumps	1.0986
over	1.0986
lazy	1.0986
dog	0.4055
is	0.0000
very	1.0986
clever	0.4055

TF-IDF Combined

Term	Document 1	Document 2	Document 3
the	0.0000	0.0000	0.0000
quick	0.1761	0.3522	0.3522
brown	0.1761	0.1761	0.3522
fox	0.4055	0.4055	0.4055
jumps	1.0986 x 1/9 = 0.1221	0.0000	0.0000
over	1.0986 x 1/9 = 0.1221	0.0000	0.0000
lazy	1.0986 x 1/9 = 0.1221	0.0000	0.0000

Thank you

FOR FURTHER QUERIES PING ME ON SKYPE GROUP