Lecture 10

-NLP is all about words, their arrangement and their meaning

-NLP is a branch of AI that tries to emulate/understand the way a human speaks to another human

-Lecture 10 is about N-gram features for text classification

N-gram

- Sequence of N-tokens, N=1,2,…
- Concept of a phrase containing N tokens or words
- Example of unigram (or 1-gram): “Operating”
- Example of bigram: “Operating system”   “Operating theater”
- Example of trigram: “Natural Language Processing”
- Some researchers use mix of unigram, bigram and trigram for their text classification projects (concatenation)

Example of bigram extraction

- Text doc 1: She sells sea shells on the sea shore
- Bigrams: (7 unique bigrams – arrange along the columns (BoW))
- she sells sea shells on the sea shore
- she sells sea shells on the sea shore
- she sells sea shells on the sea shore
- she sells sea shells on the sea shore
- she sells sea shells on the sea shore
- she sells sea shells on the sea shore
- she sells sea shells on the sea shore

BoW_features= [1 1 1 1 1 1 1]
(One-hot)

Order of bigram features (feature columns)

- she sells
- sells sea
- sea shells
- shells on
- on the
- the sea
- sea shore

Example of bigram extraction

- Text doc 2: On the sea shore, she sells.
- Bigrams:
- on the sea shore she sells
- on the sea shore she sells
- on the sea shore she sells
- on the sea shore she sells
- on the sea shore she sells

Order of 7 bigram features (feature columns)

she sells	1
sells sea	
sea shells	
shells on	
on the	1
the sea	1
sea shore	1

Example of bigram extraction

- Text doc 2: On the sea shore, she sells.
- Bigrams:
- on the sea shore she sells
- on the sea shore she sells
- on the sea shore she sells
- on the sea shore she sells [shore she is not a known bigram –so?]
- on the sea shore she sells

BoW_features= [1 0 0 0 1 1 1]
(One-hot)

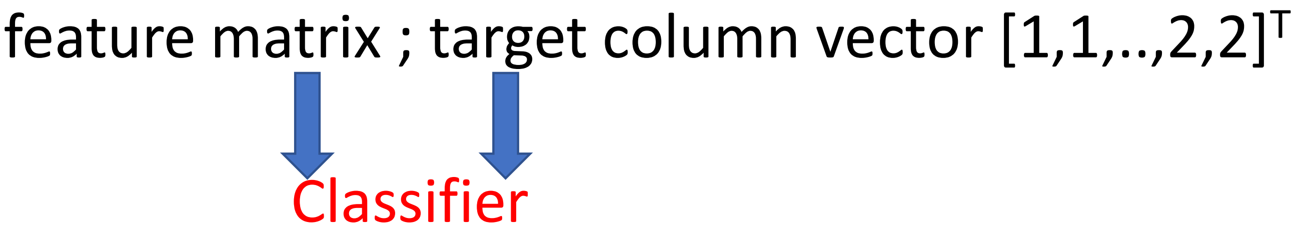
When you encounter unknown bigrams……

- Two options:

1) Add the new bigram in the feature list (8 bigram features for all documents)
2) Note that above action is possible only while training (After all training documents are processed and feature vectors are extracted, the order of features is freezed)
3) Create a feature column for <UNK> unknown bigrams that occur in the test document

Supervised learning (Training, Testing phases)

Training :



feature matrix ; target column vector $[1,1,..,2,2]^T$

Classifier

([……]T : Transpose of a vector (since it is vertically arranged))

Testing: Test doc→Test feature vector (same feature columns as in training)→Trained Classifier model→Class label of Test doc

Ex – Sentiment Analysis project where Class 1 is positive sentiment and Class 2 is negative sentiment

Ex – NER project where Class 1 is Delhi and Class 2 is Trump