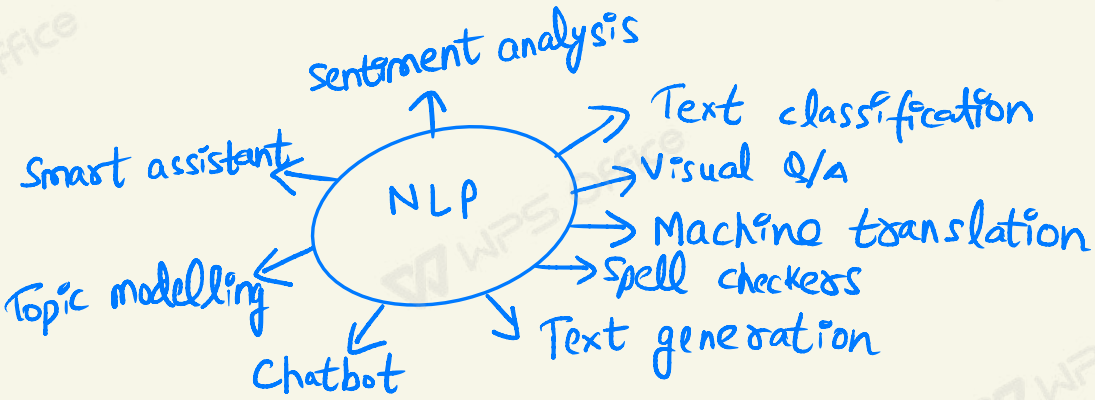
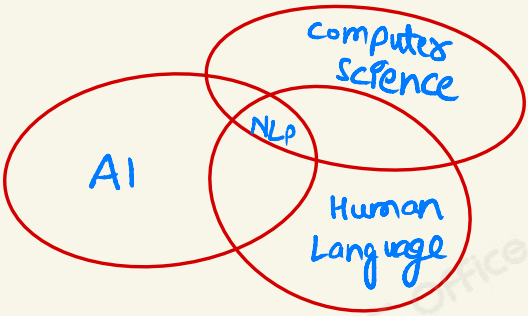
Application of NLP



Best NLP language models

1. BERT → Bidirectional Encoder representation from Transformation
2. ROBERTa → Robustly optimized BERT pre-trained approach
3. XLNET
4. Open AI's GPT 3
5. ALBERT
6. T5



→ NLP is a branch of AI which deals with communication between human language and computer that allows machine/computer to understand human language.

Most popular Libraries of NLP

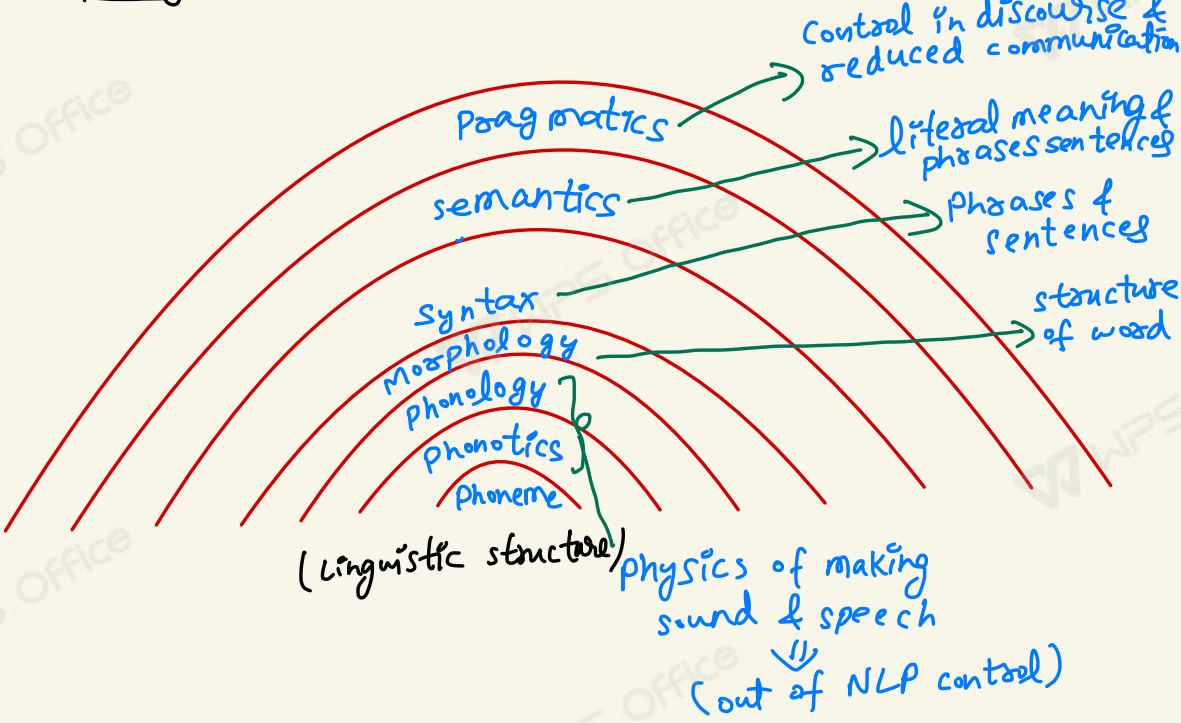
1. Hugging face (HF) Transformer
2. SpaCy
3. NLTK
4. Gensim
5. Fairseq

HW zero-shot, few-shot } self supervised, unsupervised, semi-supervised

Classification

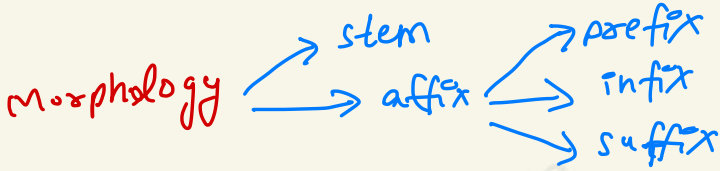
- Binary classification (2-classes)
- Multiclass classification (>2)

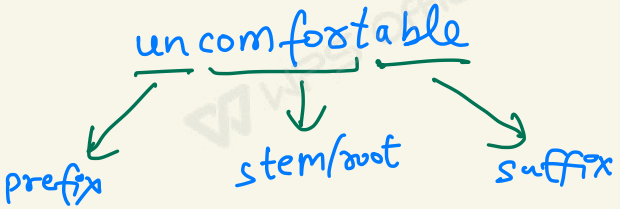
Linguistic structure



morphology → study of structure of word.

Lexicography → study of identifying words & phrases





(1) Inflection (2) Derivation (3) compounding

stem → plural, past, progresive

Ex:- jump → jumps, jumped, jumping
like → likes, liked, liking

(2) Derivation

print + er → printer
re + print → reprint
print + able → printable

(3) compounding

corpora → corpus

text data

stem + stem

house + boat → houseboat
boat + house → boathouse

(order matters)

morphological rich ⇒ order does not matter (same meaning)

Ex's

Will → modal verb

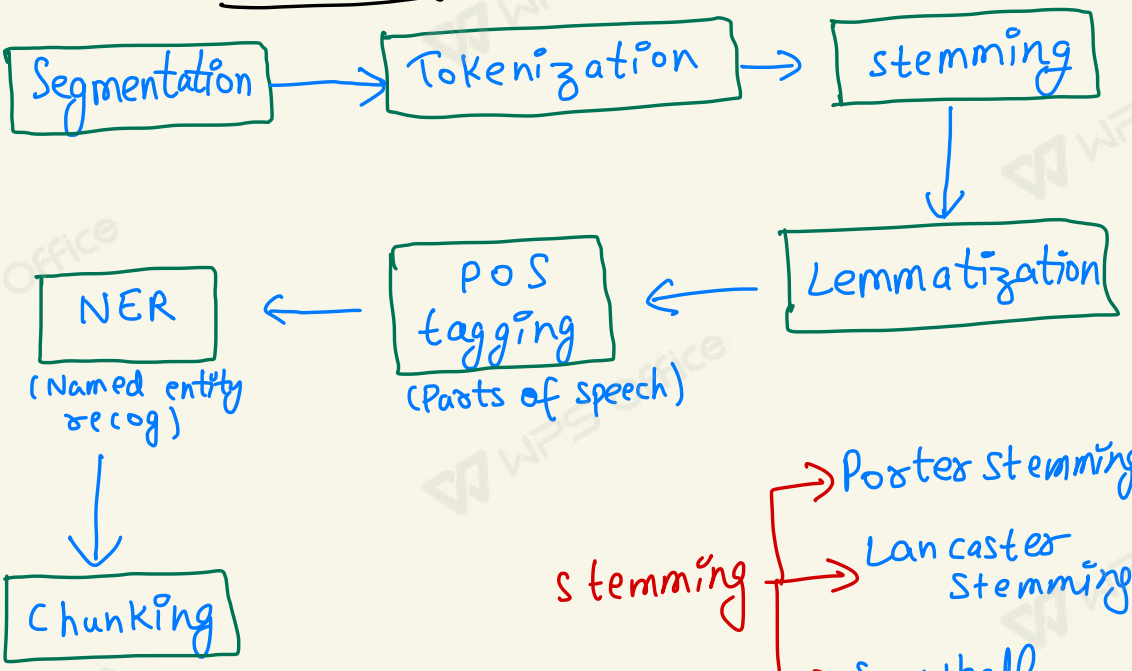
Will → person name

will → verb

Will's → person (same / diff)

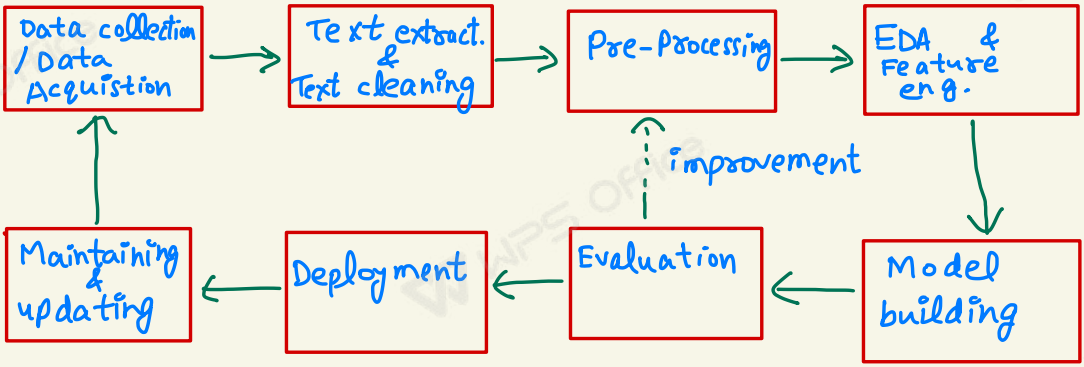
will → desire

NLP stages / NLP Pipeline



stemming
- Porter stemming
- Lancaster Stemming
- Snowball stemming

NLP PIPELINE (In widen properties) / Lifecycle



Probability Language model

↳ Joint probability
↳ conditional probability
↳ marginal probability

$P(A,B) = P(A) \cdot P(B/A)$

$P(A,B,C) = P(A)\, P(B/A)\, P(C/A,B)$

$P(A,B,C,D) = P(A) \cdot P(B/A) \cdot P(C/A,B) \cdot P(D/A,B,C)$

Markov's Assumption

→ n-gram
→ uni-gram
→ bi-gram $P(w_i / w_{i-1})$
→ tri-gram $P(w_i / w_{i-2}\, w_{i-1})$

Probability Overview: Basics of Probability Theory

tossing two coins simultaneously,

possible outcomes ⇒

$S = \{TT, TH, HT, HH\}$

1. Sample space
2. Random experiment
3. Favourable event
4. Success.
5. Random variable

Probability Distribution

Discrete

1. General Discrete
2. Binomial / Bernoulli's
3. Hyper geometric
4. Geometric
5. Poisson's

Continous

1. General continous
2. Uniform distribution
3. Exponential distribution
4. Normal / Gaussian
5. Standard normal

Random variable

- DRV → PMF (Prob mass func^n)
- CRV → pdf (Prob density func^n)

RV = x :	0	1	2
f(x) = P(x)	1/4	1/2	1/4
F(x)	1/4	1/4+1/2	1

$$\sum_x f(x) = 1$$

$$\sum_x P(x) = 1$$

(1) Expectation $E(x)$ or mean of X or Avg. of X $= \mu = \sum_x x f(x)$

(2) Variance $\quad Var(X) = \frac{1}{N}\Sigma(x-\mu)^2 = E(x^2) - E(x)^2$

(3) Standard deviation $(\sigma) = \sqrt{var(x)}$

Q. Consider the following PMF of a random variable n,

$$P(n, q) = \begin{cases} q & \text{if } x=0 \\ 1-q & \text{if } x=1 \\ 0 & \text{otherwise} \end{cases}$$

if $q = 0.4$, variance = ?

Q. A machine produces 0, 1 or 2 defective piece in a day with associated prob of 1/6, 2/3 and 1/6 respectively. Then mean value & the variance of the no. of defective pieces produced by the machine.

(a) 1, 1/3 (c) 1, 4/3

(b) 1/3, 1 (d) 4/3, 1/3

Ans 1:

X	0	1
f(n)	0.4	0.6

$E(X) = 0 \times 0.4 + 1 \times 0.6$

$= 0.6$

$\text{var}(X) = E(X^2) - E(n)^2$

$= (0 \times 0.4 + 1^2 \times 0.6) - 0.36$

$= 0.6 - 0.36$

$= 0.24$

Ans 2: $E(X) = \sum n f(n)$

$\text{var}(X) = \left(1 \times \frac{2}{3} + 1^2 \times \frac{1}{3} - 1\right) = \frac{4}{3} - 1 = 1/3$

Q. The function $P(x)$ is given by $P(x) = \frac{A}{x^\mu}$ where A and μ are constants with $\mu > 1$ and $1 \le x < \infty$ and $P(x) = 0$ for $-\infty < x < 1$ for $P(x)$ to be a probability density function the value of A should be equal to

(A) $\mu - 1$ (B) $\mu + 1$ (C) $1/\mu - 1$ (D) $1/\mu + 1$

$\Rightarrow \sum P(x) = 1$

$\Rightarrow \sum_{x=1}^{\infty} \frac{A}{x^\mu} = 1$

$\Rightarrow A\left[\frac{1}{1^\mu} + \frac{1}{2^\mu} + \frac{1}{3^\mu} + \cdots\right] = 1$

$\Rightarrow \int_{x=1}^{\infty} Ax^\mu \, dx = 1$

$\Rightarrow A \times \left[\frac{x^{-\mu+1}}{1-\mu}\right]_{x=1}^{\infty} = 1$

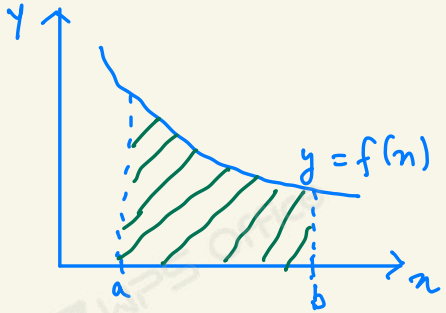
$\Rightarrow A\left[\frac{1}{x^{1-\mu}(1-\mu)}\right]_{x=1}^{\infty} = 1$

$\Rightarrow A \times \left[0 - \frac{1}{1-\mu}\right] = 1$

$\Rightarrow \frac{A}{\mu - 1} = 1$

$\Rightarrow \boxed{A = \mu - 1}$

General Continous Probability Distribution



x is a continous random variable.

$$P(x=a) = 0$$

$$P(x=b) = 0$$

$$P(a < x < b) = \int_a^b f(x)\,dx$$

(i) $x \to$ c.r.v then, probability at a single point $= 0$

$P(x=a) = 0$.

(ii) $x \to$ c.r.v. $\Rightarrow$

$$P(a \le x \le b) = \int_a^b f(x)\,dx$$

(iii) $P(-\infty < x < \infty)$ = Area under the density func^n curve.

$$= \int_{-\infty}^{\infty} f(x)\,dx$$

(iv) $f(x) \ge 0$, always positive

Q. If x is a c.r.v. having pdf is given by

$$f(x) = \begin{cases} cx^2, & 0 \le x < 1 \\ cx, & 1 \le x < 2 \\ 0, & \text{otherwise} \end{cases}$$

(i) Find c
(ii) $P(1/2 < x < 3/2)$
(iii) $P(x < 3/2)$
(iv) $P(x \ge 1/2)$

(i) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

$$\Rightarrow \int_0^1 cx^2 \cdot dx + \int_1^2 cx \cdot dx = 1$$

$$\Rightarrow \frac{cx^3}{3}\Big|_0^1 + \frac{cx^2}{2}\Big|_1^2 = 1$$

$$\Rightarrow \frac{c}{3} + \left[c \times 2 - \frac{c}{2} \right] = 1$$

$$\Rightarrow \frac{c}{3} + \frac{3c}{2} = 1$$

$$\Rightarrow \frac{2c + 9c}{6} = 1$$

$$\Rightarrow c = 6/11$$

(ii) $P(x) = \int_{1/2}^{1} cx^2 \cdot dx + \int_{1}^{3/2} cx \cdot dx$

$= \frac{cx^3}{3}\Big|_{1/2}^{1} + \frac{cx^2}{2}\Big|_{1}^{3/2}$

$= \frac{c}{3} - \frac{c}{24} + \frac{c \times 9}{8} - \frac{c}{2}$

$= \frac{6}{11}\left[\frac{1}{3} - \frac{1}{24} + \frac{9}{8} - \frac{1}{2}\right]$

$= \frac{6}{11} \times \frac{8 - 1 + 27 - 12}{24}$

$= \frac{\cancel{6}}{\cancel{11}} \times \frac{\cancel{22}^{2}}{\cancel{24}_{4}} = 1/2$

(iii) $P = \int_{0}^{1} cx^2 \cdot dx + \int_{1}^{3/2} cx \cdot dx$

$= \frac{cx^3}{3}\Big|_{0}^{1} + \frac{cx^2}{2}\Big|_{1}^{3/2}$

$= \frac{c}{3} + \frac{c \times 9}{8} - \frac{c}{2}$

$= \frac{6}{11} \times \frac{8 + 27 - 12}{24}$

$= \frac{6}{11} \times \frac{23}{\cancel{24}_{4}} = \frac{23}{44}$

(iv) $P(x \geq 1/2) = \int_{1/2}^{1} cx^2\,dx + \int_{1}^{\infty} cx\,dx$

$= \frac{cx^3}{3}\Big|_{1/2}^{1} + \frac{cx^2}{2}\Big|_{1}^{2}$

$= \frac{c}{3} - \frac{c}{24} + \frac{c \cdot 4}{2} - \frac{c}{2}$

$= \frac{6}{11}\left[\frac{1}{3} - \frac{1}{24} + \frac{4}{2} - \frac{1}{2}\right]$

$= \frac{6}{11} \times \frac{8 - 1 + 36}{24}$

$= \frac{43}{4} \times \frac{1}{11}$

Q. For the func^n $f(x) = a + bx$, $0 < x < 1$ to be a valid p.d.f. which one of the following statement is true.

(a) $a = 1, b = 4$ (b) $a = 0.5, b = 1$

(c) $a = 0, b = 1$ (d) $a = 1, b = -1$

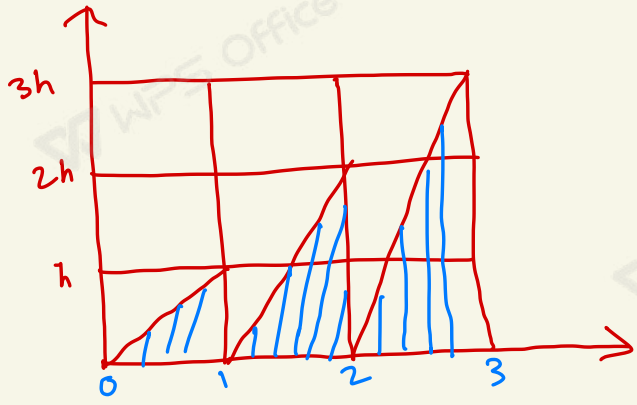
$\Rightarrow \int_0^1 (a + bx)\,dx = 1$

$\Rightarrow ax + \frac{bx^2}{2}\Big|_0^1 = 1$

$$a + \frac{b}{2} = 1$$

(b)

Q. Find the value of h,



$$\Rightarrow A_1 + A_2 + A_3 = 1$$

$$\Rightarrow \frac{1}{2} \times 1 \times h + \frac{1}{2} \times 1 \times 2h + \frac{1}{2} \times 1 \times 3h = 1$$

$$\Rightarrow h = 1/3$$

Q.

$$F(x) = \begin{cases} 0 & , -\infty < x < 0 \\ \frac{2}{11}x^3 & , 0 \le x < 1 \\ \frac{3x^2 - 1}{11} & , 1 \le x < 2 \\ 1 & , 2 \le x < \infty \end{cases}$$

(i) Find $P(\frac{1}{2} < x < 3/2)$

(ii) $P(x \geq 1/2)$

(iii) $P(x < 3/2)$

Q.1 If X is uniformly distributed in (0,10) then find

(i) $f(x)$
(ii) mean, variance, std deviation
(iii) $P(2<X<6)$
(iv) $P(0<x<5)$
(v) $P(x \leq 3)$
(vi) $P(x>8)$

Q.2 If a random variable is uniformly distributed with mean 1 and variance $1/3$, then $P(X<1/2) = ?$

Q.3 The P.d.f of a random variable X is

$f(x) = \frac{x}{4}(4-x^2)$ for $0 \leq x \leq 2$

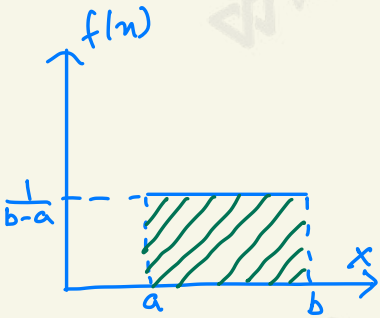
$= 0$, otherwise

Find the mean.

Uniform Density

If X is uniformly distributed continuous random variable, then the probability density function in the finite interval $[a,b]$ is given as

$$f(x) = \begin{cases} \frac{1}{b-a} & , a \leq x \leq b \\ 0 & , \text{otherwise} \end{cases}$$



$$f(x) = \begin{cases} k & , a \leq x \leq b \\ 0 & , \text{otherwise} \end{cases}$$

Properties of Pdf

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)\,dx$

Ans 1 :- (i) $f(x) = \frac{1}{b-a} = \frac{1}{10-0} = \frac{1}{10}$

$$\therefore f(x) = \begin{cases} 1/10 & , 0 < x < 10 \\ 0 & , \text{else} \end{cases}$$

(ii) $\text{mean} = E(x) = \mu = \int_{-\infty}^{\infty} x f(x)\,dx$

$$= \int_a^b \frac{1}{b-a} \cdot x\,dx$$

$$= \frac{1}{2(b-a)} [x^2]_a^b$$

$$= \frac{1}{2(b-a)} b^2 - a^2$$

$$\text{mean} = \frac{b+a}{2}$$

$$\therefore \text{mean} = \frac{10+0}{2} = 5$$

$$\text{variance} = E(x^2) - (E(x))^2$$

$$E(x^2) = \int_a^b \frac{1}{b-a} x^2\,dx$$

$$= \left[\frac{x^3}{3(b-a)}\right]_a^b = \frac{b^3 - a^3}{3(b-a)}$$

$$E(x) = \left(\frac{b+a}{2}\right)$$

$$E(x^2) = \frac{(b-a)(b^2+a^2+ab)}{3(b-a)}$$

$$(E(x))^2 = \frac{(b+a)^2}{4} = \frac{a^2+b^2+2ab}{4}$$

$$\text{variance} = \frac{b^2+a^2+ab}{3} - \frac{a^2+b^2+2ab}{4}$$

$$= \frac{4b^2+4a^2+4ab-3a^2-3b^2-6ab}{12}$$

$$= \frac{b^2+a^2-2ab}{12}$$

$$= \frac{(b-a)^2}{12} = \frac{100}{12}$$

$$\text{std deviation} = \frac{10}{\sqrt{12}}$$

Ans 2: $\text{mean} = 1 = \frac{b+a}{2}$

$b+a = 2$

$\text{variance} = 1/3$

$\frac{(b-a)^2}{12} = \frac{1}{3}$

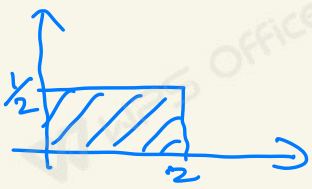
$\frac{(b-a)^2}{4} = 1$

$b-a = 2$

$b = 2$

$a = 0$

$P(x < 1/2) = \frac{1}{4}$



Ans 3 :- $f(x) = \frac{x}{4}(4-x^2)$ for $0 \le x \le 2$

$= 0$, otherwise

$\Rightarrow \int x \cdot \frac{x}{4}(4-x^2)\, dx$

$$E(X) = \frac{1}{2\pi}\int_0^{2\pi} \cos(2\pi t_1 - 2\pi t_2) - \cos(2\pi t_1 + 2\pi t_2 + 2\phi) \cdot d\phi$$

$$= \frac{1}{2\pi}\Big[\sin(2\pi t_1 - 2\pi t_2) - \sin(2\pi t_1 + 2\pi t_2 + 2\phi)\Big]_0^{2\pi}$$

$$= \frac{1}{2\pi}$$

continuous Pd

$\mu = \frac{b+a}{2}$

$\text{variance} = \frac{(b-a)^2}{12}$

$\sigma = \frac{b-a}{\sqrt{12}}$

$\text{Pdf} \Rightarrow f(x) = \begin{cases} \frac{1}{b-a}, & a<x<b \\ 0, & \text{otherwise} \end{cases}$

Exponential Pd

$f(x) = \begin{cases} \alpha e^{-\alpha x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$

$\mu = E(x) = \int_{-\infty}^{\infty} x f(x)\, dx$

$= \int_{0}^{\infty} x \cdot \alpha e^{-\alpha x} dx$

$= x \cdot \frac{\alpha e^{-\alpha x}}{-\alpha} + \int \frac{\alpha e^{-\alpha x}}{\alpha} dx$

$= -x e^{-\alpha x} - \frac{e^{-\alpha x}}{\alpha} \Big|_0^{\infty}$

$= 0 - \left[0 - \frac{1}{\alpha}\right]$

$\boxed{\mu = 1/\alpha}$

$\text{Variance} = E(x^2) - E(x)^2$

$= \alpha \int_0^{\infty} x^2 e^{-\alpha x} dx$

$= \alpha \left[x^2 \cdot \frac{e^{-\alpha x}}{-\alpha} + \int 2x \cdot \frac{e^{-\alpha x}}{\alpha} dx \right]$

$= \alpha \left[x^2 \frac{e^{-\alpha x}}{-\alpha} + \left[2x \times \frac{e^{-\alpha x}}{-\alpha^2} - \int \frac{2x e^{-\alpha x}}{-\alpha^2} dx \right] \right]$

$= -x^2 e^{-\alpha x} - \frac{2x e^{-\alpha x}}{\alpha} + \frac{2 e^{-\alpha x}}{\alpha^2} \Big|_0^{\infty}$

$\frac{2}{\alpha^2}$

$$\text{Variance} = \frac{2}{\alpha^2} - \frac{1}{\alpha^2} = \frac{1}{\alpha^2}$$

$$\sigma = 1/\alpha$$

Q. If the call duration is ed. with $\alpha = 1/10$ then find

(i) call duration exceeds 7mins
(ii) " between 3 & 5 mins
(iii) " less than 8
(iv) greater than avg of call duration.

=>
$$f(x) = \begin{cases} \frac{1}{10} e^{-\frac{x}{10}} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$P(x > 7) = \int_7^\infty \frac{1}{10} e^{-\frac{x}{10}} \cdot dx$$

$$= \frac{1}{10} \frac{e^{-x/10}}{-1/10} \Big|_7^\infty$$

$$= e^{7/10}$$

(iv)
$$\int_3^5 \frac{1}{10} e^{-\frac{x}{10}} \cdot dx$$

$$= -e^{-x} \Big|_3^5 = -e^{-5/10} + e^{-3/10}$$

(iii) $\int_0^8 \frac{1}{10} e^{-x/10}\, dx$

$= -e^{-\frac{x}{10}} \Big|_0^8 = -e^{-8/10} - 1$

(iv) $\alpha = 1/10$, $\mu = 10$

$P(x > 10)$

Q Let Z be ERV - Mean = 1

$P(Z > 2 \mid Z > 1)$

$\Rightarrow \frac{1}{\alpha} = 1 \Rightarrow \alpha = 1$

$f(x) = \begin{cases} 1\, e^{-x} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$

$P(Z > 2) = \int_2^\infty e^{-x} \cdot dx$

$= -e^{-x} \Big|_2^\infty$

$= +e^{-2}$

$$P(Z>1) = \int_1^\infty e^{-x}\,dx$$

$$= -e^{-x}\Big|_1^\infty = e^{-1}$$

$$P(Z>2 \mid Z>1) = \frac{e^{-2}}{e^{-1}} = e^{-1}$$

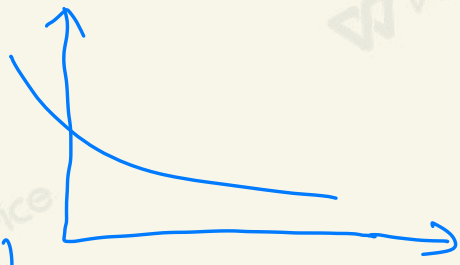
Q. Let X_1 & X_2 be two independent ERV. with mean = 0.5 & 0.25 respectively then $Y = \min(X_1, X_2)$. what

$$\alpha_1 = \frac{1}{0.5} = 2$$

$$\alpha_2 = \frac{1}{0.25} = 4$$

$$f(x_1) = \{ 2e^{-2x}$$

$$f(x_2) = \{ 4e^{-4x}$$



$$= \min(2e^{-2x}, 4e^{-4x})$$

=

$$Var(X+Y) = Var(X) + Var(Y) + 2cov(X,Y)$$

if X & Y are independent variable,

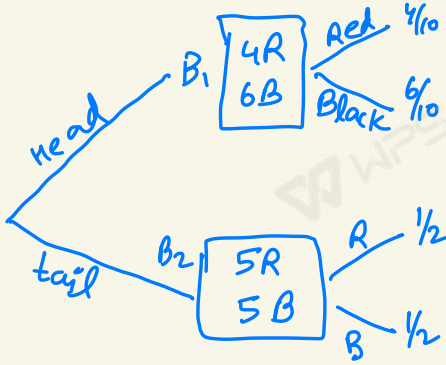
$$V(X+Y) = Var(X) + Var(Y)$$

$$\text{corelation coeff} = \gamma = \frac{cov(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}}$$

Q. consider two boxes box1 & box2. Box 1 contains 4 red and 6 black balls whereas Box2 contains 5 red and 5 black balls. Now, a coin is tossed, if head occurs then one ball is randomly drawn from box1. whereas if tail occurs then one ball from Box2.

(i) Find the probability of getting a red ball.

(ii) if the ball obtained is red, what is the prob that it comes from box1.



$\Rightarrow P(B_1) = 1/2 = P(B_2)$

$P(Red) = P(Red \cap B_1) + P(Red \cap B_2)$

$P(Red) = P(A)\cdot P(E|A) + P(B)\cdot P(E|B)$

$= \frac{1}{2} \times \frac{4}{10} + \frac{1}{2} \times \frac{1}{2}$

$= 9/20$

(ii) $\frac{\text{fav. path}}{\text{total path}} = P(B_1|Red) = \frac{P(B_1) \cap P(Red)}{P(Red)}$

$= \frac{\frac{1}{2} \times 4/10}{9/20} = 4/9$

Types of Events :-

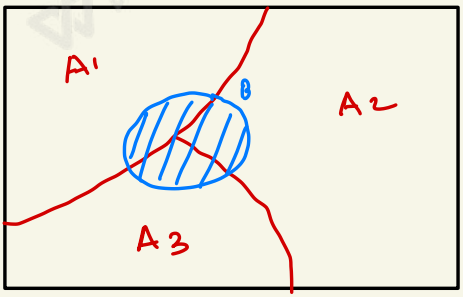
1) Mutually dependent
2) Mutually Independent $(P(A \cap B) = P(A) \cdot P(B))$
3) Mutually Exclusive (either head or tail)
4) Collectively exhaustive $(P(A) + P(B) + P(C) = 1)$

$J = M \cdot C$

- Joint prob
- marginal prob (independent)
- conditional (dependent)

$$P(A \cap B) = P(A) \cdot P(B|A)$$

→ If $P(A \cap B) = 0$, then A & B are mutually exclusive.



$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$
$$= P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2) + P(A_3) \cdot P(B|A_3)$$

$$P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)$$

Total probability theorem

Baye's theorem

$$P(A_2|B) = \frac{P(A_2 \cap B)}{P(B)} = \frac{P(A_2) \cdot P(B|A_2)}{P(B)}$$

$$P(A_2|B) = \frac{P(A_2) \cdot P(B|A_2)}{P(B)}$$

Q. In a bolt factory, machines A, B & C manufacture 25%, 35% & 40% respectively of the total output There is a chance of 5%, 9% and 2% resp are defective. A bolt is drawn at random.

(i) Prob that it is a defective bolt 0.0345

(ii) Prob that defective is drawn from A. 0.3623

$$\therefore P(def) = P(def \cap A) + P(def \cap B) + P(def \cap C)$$

$$= \frac{25}{100} \times \frac{5}{100} + \frac{35}{100} \times \frac{9}{100} + \frac{40}{100} \times \frac{2}{100}$$

$$= \frac{125 + 315 + 80}{10000}$$

$$= \frac{520}{10000} =$$

Q. Consider the following corpus of 4 sentences.
<s> three friends Aman, akbar and anthony are reading book. </s>
<s> aman is reading malgudi days </s>
<s> akbar is reading a detective book </s>
<s> anthony is reading a book by nk narayan </s>
Assume a bigram language model. Calculate P(<s> aman is reading a book </s>)

Q. There are 3 bags B_1, B_2 & B_3. The bag B_1 contains 5 green & 5 red balls, bag B_2 contains 3 red & 5 green balls and bag B_3 contains 5 red & 3 green balls. Bags B_1, B_2 & B_3 have probabilities 3/10, 3/10 & 4/10 respectively if being chosen. A bag is selected at random and a ball is chosen at random from bag.

(A) Find the probability that the chosen ball is green, given that the selected bag is B_3.

(B) " " " " is green.

(c) Find the probability that the selected bag is B_3, given that the chosen ball is green.

(d) Find the prob. that the selected Bag is B_3 & given that the chosen ball is green.

Ans 2 :-
(B) $P(\text{green}) = \frac{3}{10} \times \frac{5}{10} + \frac{3}{10} \times \frac{5}{8} + \frac{4}{10} \times \frac{3}{8}$

$= 0.15 + \frac{15}{80} + \frac{12}{80}$

$= 0.15 + 0.1875 + 0.15$

$= 0.4875$

(A| =

$$\frac{4/10 \times 3/8}{4/10}$$

$$= 3/8$$

(c) $p = \frac{4/10}{0.4875} = 0.8205$

(d) $p = \frac{4}{10} \times 0.4875$

$= 0.195$

$P(B_3 \cap \text{Green}) = P(B_3) \cdot P(\text{green}|B_3)$

$= 3/20$

Ans 1 :-

$$p = \frac{3+6+8+4+4}{24} = \frac{25}{24}$$

Q. Consider the following corpus of 3 sentences what is the total count of unique bigrams for which the likelihood will be estimated assume we do not perform any preprocessing. consider the beginning & end of token as <s> & </s>

-> <s> Julia is visiting the museum </s>
<s> Julia, grover & natasha are friends </s>
<s> Zoe & natasha will meet Julia in the museum </s>

(a) 23 (b) 20 (c) 16 (d) 18

Function words vs content words

stop words
or
closed class words
a, an, the, to, is, of, ---
prepositions – in, on, of, by
determiner – a, an, the
pronouns – I, he, his, him, me, mine

info / topic
open class words
↳ we keep on adding new words.
noun
verb
adjective
adverbs

Type Token Ratio (TTR)

$$TTR = \frac{types}{tokens} = \frac{\text{no. of unique words}}{\text{total words}}$$

Ex:- will will will

TTR = 1/3

* If TTR is high, then new words will be found more.

Q. In a corpus, it was found that the word with rank 4th has a frequency of 600. What can be the best guess for the rank of a word with freq. 300.

$\Rightarrow 1600 \times 4 = 300 \times n$

$n = 8$

Q. In the sentence, "The only thing we have to fear is the fear itself", Find the TTR. $\Rightarrow \frac{9}{11}$

Q. Let the rank of 2 words w_1 & w_2 in a corpus be 1600 and 400 respectively. Let m_1 & m_2 represent the no. of meanings of w_1 and w_2 respectively. The ratio of $m_1 : m_2$ would tentatively —

$m_1 \times \sqrt{1600} = m_2 \times \sqrt{400}$

$\Rightarrow 1/2$

Q. which are true :-

T (a) Ambiguity can appear in Tokenization steps.

F (b) Ambiguity will not appear in sentence segmentation step.

T (c) Function used is generally more frequent in a text than any any content word.

T (d) output of lemmatization are always real words

Zipf's Law

$f \propto \frac{1}{r}$

f: freq of word
r: rank of word in desc ord

$\therefore f \cdot r = k = \text{constant}$

P_r (prob of word of rank r) $= \frac{f}{N} = \frac{k}{r \cdot N}$

$= \frac{A}{r}$

$m \propto \sqrt{f}$

$m \propto \frac{1}{\sqrt{r}}$

$f \propto \frac{1}{l}$

m : no. of meanings

r : rank

l : length of a word

Heap's Law

$|V| = KN^{\beta}$, $|V|$ = size of vocabulary
N = number of tokens

Q. what is the size of unique words in a document where total no. of words is 12000, $K = 3.71$ & $\beta = 0.69$.

$$|V| = 3.71 \times (12000)^{0.69}$$
$$= 2421$$

Q. If the first corpus has TTR = 0.085 & second corpus has TTR = 0.78. which of the following are F.

F (i) 1st corpus has more tendency to have unique words.

T (ii) 2nd " " " " " "

F (iii) TTR values can have sometimes > 1.

T (iv) ↑ TTR indicates ↑ degree of lexical variance & vice versa.

1) Ambiguity in lexicography

2) Issues in Tokenization

10-12 slides

Conversion of text into vectors

(vectorization)

for feature extraction or Information retrieval

1. BoW → Bag of words model
2. Binary BoW
3. TF-IDF (Term frequency-Inverse doc freq)

ML → label encoding & OHE (one hot encoding)

Corpus

S_1: He is an awesome boy and an awesome dancer too.

S_2: She is also an awesome girl

S_3: Both the girl and the boy are awesome.

S_4: It was a good movie with awesome acting.

After removal of stopwatch

Vocabulary		freq
awesome	f_1	5
boy	f_2	2
dance	f_4	1
girl	f_3	2
good	f_5	1
movie	f_6	1
acting	f_7	1

Text pre-process:

1. Stopwords removal
2. lowering of the case
3. stemming & lemmatization
4. Removal of punctuation sym.
5. handling of negation

Vector for 1st sentence — awesome boy awesom dancer

f_1	f_2	f_3	f_4	f_5	f_6	f_7
awesome	boy	girl	dancer	good	movie	acting
[2	1	0	1	0	0	0]

Vectors → fixed size

Q. S_1: The pizza was good
S_2: The pizza was not good

→

	good	the	pizza	was	not
S_1 :	[1	1	1	1	0]
S_2 :	[1	1	1	1	1]

$$TF(t,d) = \frac{\text{total no. of times 't' present in doc. d}}{\text{total no. of tokens in the given doc d}}$$

(Term freq.)

$$IDF(t) = \log\left(\frac{\text{Total no. of documents}}{\text{no. of documents containing the term t}}\right)$$

(Inverse document freq.)

$$TF\text{-}IDF = TF(t,d) \times IDF(t)$$

↓ score

Q. d_1: He is a good boy
d_2: She is a good girl
d_3: Both the boy and the girl are good
Calculate TF-IDF score for the above corpus.

→ d_1 : good boy
d_2 : good girl
d_3 : boy girl good

	d_1	d_2	d_3
good:	$\frac{1}{2}$	$1/2$	$1/3$
boy:	$1/2$	0	$1/3$
girl	0	$1/2$	$1/3$

(TF(t,d))

good: $\log \frac{3}{3} = 0$

boy: $\log \frac{3}{2}$

girl: $\log 3/2$

(IDF(t))

	good	boy	girl
d_1:	0	$\frac{1}{2} \times \log \frac{3}{2}$	0
d_2:	0	0	$\frac{1}{2} \log \frac{3}{2}$
d_3:	0	$\frac{1}{3} \log \frac{3}{2}$	$\frac{1}{3} \log \frac{3}{2}$

(TF-IDF | $(d_i \times IDF_i)$)

Q. From a website we got 3 reviews of a movie:
R1: This movie is very scary and long.
R2: This movie is not scary and is slow.
R3: This movie is spooky and good.

⇒

	d_1	d_2	d_3	IDF
This:	$1/7$	$1/8$	$1/6$	$\log 3/3 = 0$
movie:	$1/7$	$1/8$	$1/6$	$\log 3/3 = 0$
is:	$1/7$	$2/8$	$1/6$	$\log 3/3 = 0$
very:	$1/7$	0	0	$\log 3/1$
scary:	$1/7$	$1/8$	0	$\log 3/2$
and:	$1/7$	$1/8$	$1/6$	$\log 3/3 = 0$
long:	$1/7$	0	0	$\log 3/1$
not:	0	$1/8$	0	$\log 3/1$
slow:	0	$1/8$	0	$\log 3/1$
spooky:	0	0	$1/6$	$\log 3/1$
good:	0	0	$1/6$	$\log 3/1$

	is	This	movie	and	scary	very	long	spooky	slow	not	good
R_1:	0	0	0	0	$\frac{1}{7}\log\frac{3}{2}$	$\frac{1}{7}\log 3$	$\frac{1}{7}\log 3$	0	0	0	0
R_2:	0	0	0	0	$\frac{1}{8}\log\frac{3}{2}$	0	0	0	$\frac{1}{8}\log 3$	$\frac{1}{8}\log 3$	0
R_3:	0	0	0	0	0	0	0	$\frac{1}{6}\log 3$	0	0	$\frac{1}{6}\log 3$

Spelling correction - Edit Distance

I am writing an email on behaf of KIIT.

incorrect

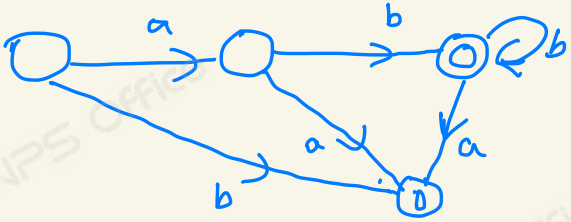
↳ correct one
- behalf
- behave
- behaviour

min. distance
⇓
min no. of operations
↳ Insertion
↳ deletion
↳ substitution

} Levenshtein distance

$\{a, b\}$

abb^* --



N	9	8	9	10	11	12	11	10	9	8
O	8	7	8	9	10	11	10	9	8	9
I	7	6	7	8	9	10	9	8	9	10
T	6	5	6	7	8	9	8	9	10	11
H	5	4	5	6	7	8	9	10	11	10
E	4	3	4	5	6	7	8	9	10	9
T	3	4	5	6	7	8	7	8	9	8
N	2	3	4	5	6	7	8	7	8	7
I	1	2	3	4	5	6	7	6	7	8
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

$$D(i,j) = \min \begin{cases} D(i-1,j)+1 \\ D(i,j-1)+1 \\ D(i-1,j-1)+n \end{cases}$$

$n = 2$ if $x(i) \neq y(j)$

$= 0$ if $x(i) = y(j)$

Q. "The boy put the keys on the table". Assign the POS tag for each word.

=>
- The → det
- boy → noun
- put → verb
- the → det
- keys → noun
- on → pre
- the → det
- table → noun

Q. "The grand jury commented on a number of other topics".

=>
- The → DT (determiner)
- grand → JJ (adjective)
- jury → NN (noun)
- commented → VBD (verb)
- on → pre - IN
- a → det - DT
- number → noun - NN
- of → prep - IN
- other → JJ (adjective)
- topics → NNS (singular noun)

Q. "I need a flight from Atlanta".

⇒ I → PRP (personal pronoun)
need → VB (verb)
a → DT (determiner)
flight → NN
from → IN
Atlanta → NNP

$P(VB \mid TO) \times P(NR \mid VB) \times P(race \mid VB)$

$P(VB \mid TO)$: state transition prob.
$P(NR \mid VB)$: STP
$P(race \mid VB)$: emmision prob.

Q. (1) Sun rises in the east. $P=1, I=0$
(2) India got Independence in 1947. $P=1, I=0$
(3) You will get 5 emails in next one year. $P=0.04$ $I\uparrow$
(4) The prime minister will come to your home tomorrow. $P=0.01$ $I\uparrow$
(5) It will snow in Delhi in June. $P=0.002$ $I\uparrow$
(6) You will not get a holiday next Sunday. $P=0.04$ $I\uparrow$
(7) The dog will bark. $P\uparrow$ $I\downarrow$

$$P \propto \frac{1}{I}$$

I: Information content

$$p(x_i) \propto \frac{1}{I(x_i)}$$

$$p(x_i) = f\left(\frac{1}{I(x_i)}\right)$$

$$p(x_i) = \log \frac{1}{I(x_i)}$$

$$= \log (I(x_i))^{-1}$$

$$\boxed{p(x_i) = -\log I(x_i)}$$

$$P(x_i)P(y_i) = I(x_i) + I(y_i)$$

$$\text{Entropy}(H) = \sum_{i=1}^{n} P(x_i) I(x_i)$$

$$= \sum_{i=1}^{n} P(x_i) \cdot \log \frac{1}{P(x_i)}$$

$$H = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

$\sum P(x_i) \log \frac{1}{P(x_i)}$

$L_{min} = H$

$L_{avg} = \sum P_i L_i$

$$P(X, Y) = P(X) \cdot P(Y|X)$$

↓ Independent

$$\therefore H(X, Y) = H(X) + H(Y|X)$$

↓ Independent

chain rule of entropy