Datascience

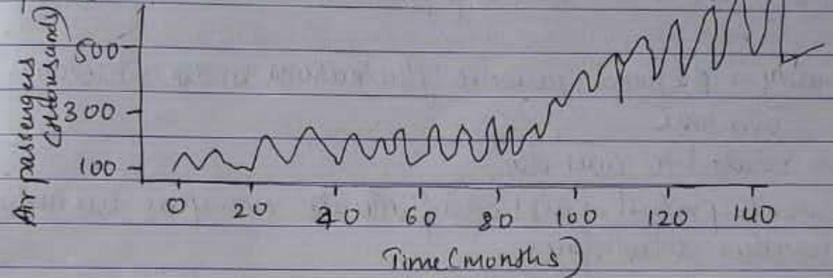
Unit4:-

* overview of time Series analysis

→ attempts to model the underlying structure of observations taken over time

→ A time series, denoted Y=a+bX

monthly number of international airline passengers over 12-year period



→ goals of time series analysis:
- Identify and model the structure of the time Series
- Forecast future values in time series

Examples:- ① Retail sales:- for various product lines, a clothing retailer is looking to forecast future monthly sales. These forecasts need to account for the seasonal aspects of customers purchasing decisions.

Forex:- in northeen hemisphere, sweater sales are typhically brisk in the fall season, and swimsuit sales are highest during the late spring and early summer.

Thus an appropriate time series model needs to account for fluating demand over the calendar year.

- stock trading:- some high frequency stock traders utilize a technique → pairs trading.
an identified strong positive corelation b/w the prices of two stocks is used to detect a market opportunity. suppose the stock prices of comp A and comp B consistently move together. TSA can be applied to the diff of these companies SP overtime

* Box - Jenkins Methodolgogy

→ Trend → seasonality → cyclic → Random

② ↓

→ longterm movement in a TS.

→ Indicate whether obs valuce are increasing or decreasing overtime

→ Examples of trends are a steady increase in sales month over month or an annual decline of fatalities due to car accidents.

② Seasonality → the fixed, periodic fluctuations in the observations over time

↓ often related to calendar.

Ex:- monthy retail sales can fluctuate over year due to the weather & holidays

③ Cyclic:- periodic fluc~~ation~~tuation, but one that is not as fixed as in the seasonality components.

Ex:- retail sales are influenced by the general state of the economy.

④ Random →

George Box and Gwilym J$\overset{n}{e}$kins → Box-Jenkins methodolgy for time series analysis involves the following 3 main steps:

1) condition data and select a model.

- Identify and account for trendys or seasonality in the time series
- examine the remaining time series and determine a suitable model

2) Estimate the model parameters

3) Assess the model and return to step1, if necessary

* ARIMA model → Auto Regressionve Integrated Moving Average

→ a class of statistical model for analyzing and forecasting time series data

→ Data shows evidence of non-stationarity

→ A random variable that is time series is stationary if its statistical properties are all constant over time

→ A stationary series has no trend, its variations around its mean have a constant amplitude, and it wiggles in a consistent fashion

→ The latter condition means that its auto correlation remain constant over time

∴ Power spectrum remains constant

→ A random variable of this form can be viewed as a combination of signal and noise

→ An arima model can be viewed as a "filter" that tries to separate the signal from the noise, and the signal is then extrapolated into the future to obtain forecasts

Q what is ARIMA forecasting eq^n for a stationary time series?

→ A linear equation in which the predictors consists of Lags of the dependent variable and /or lags of the forecast error

Predicted value of Y = a constant
a weighted sum of one or more recent values of Y
" " " " " " of the error

AR → Auto regressive - uses the dependent relationship b/w an
p ↓ observation and some number of lagged observation
Lag order

I → Integrated → The use of differencing of raw observations
d ↓ sub ~~one~~ ob from another ob ~~in order to make~~
degree of differing ~~time~~ previous time step in order to make time series stationary

MA → moving average → uses the dependency b/w
q ↓ an observation and residual errors from a moving average model applied to lagged observation
order of moving average

* Text Analysis :- called Text analytics

→ refers to representation, processing and modeling of textual data to ~~dove use~~ derive useful insights

→ ~~The~~ An important component of text analysis is Text mining

The process of discovering relationships and interesting patterns in large text collections

Text Anlysis steps :-
- ① Parsing
- ② Search and retrieval
- ③ Retrieval text mining

① Parsing :- process that takes unstructured text and imposes a structure for further analysis

→ The unstructured text could be a plain text file, a web log, an Extensible Markup lang (XML), HTML file or a word document

→ Parsing deconstructs the provided texts and renders it in a more structured way for the subsequent steps

② Search and Retrieval :- is the identification of the documents in a corpus that contain search items such as specific words, phrases or entities like people or organisations.

→ These search items are → key terms

→ Now are used by web engineers.

③ Text-mining :- uses the terms and indexes produced by the prior two steps to discover meaningful insights pertaining to domains or problems of interest.

POS (Parts of speech) → Tagging, Lemmatization, stemming

The goal of POS tagging is to built a model whose input is a sentence, such as

1. he saw a fox

and whose output is a tag sequence.

Each tag marks the POS for the corresponding word such as;

* Collecting raw text:-

# determining sentiments

The company's success depends on customers - ~~One~~ ~~to check~~ customers feedback.

Process of computationally identifying and categorizing opinions from piece of text, and determine whether the writer's attitude towards a particular topic or the product, is positive, negative or neutral.

How does it work

Step 1 :- Tokenization -> dividing a paragh into different sit of statements words

The movie was great!

↓ Tokenization

- The
- Movie
- was
- great
- !

Step2 :- Cleaning the data -> remove the special characters & or the words which do not add anything to the analytical part:

- The
- Movie
- was
- great

left with 4 words

~~!~~

Step 3:- Removing stop words -> do not add any value to the analytic result

~~. The~~ ~~. was~~

. Movie . great

1. PRP VBD DT NN
a/c to the Penn Treebank POS tags. ∴ the 4 words are mapped to pronoun (personal), verb (past tense), determiner, and noun (singular), respectively.

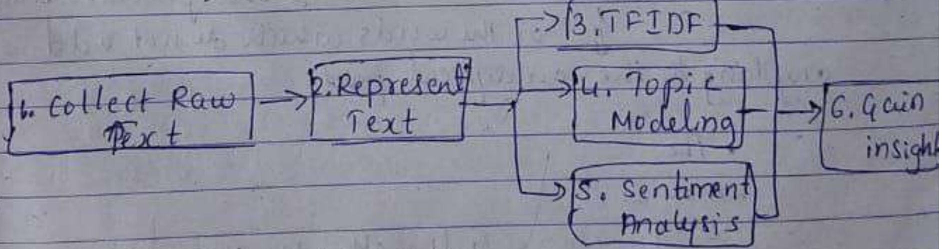
Both lemmatization & stemming are techniques to reduce the number of dimensions and reduce inflections or variants forms to the base form to more accurately measure the number of times each word appears.

1. obesity causes many problems
the o/p of Lemmatization would be:
1. obesity cause many problem

Porter's stemming algorith
1. obesity causes many problems
o/p of Porter's stemming algorithm is:

1. obes caus mani problem

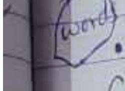
* Example
Consider company ACME, maker of two products bphone and bEbook. ACME is in strong competition with other companies that manufacture & sell similar products.



stęp4 !- classification
↓
whether it is a positive word /negative word
or neutral word

positive : sentiment score +1
negative :- −1
neutral :- 0

Apply supervised algorithm for classification



. Train your model with bag of words or Lexicons, (dict of pre classified set of words) and test it on the analysis statement

. More the accuracy score better will be the classification

Movie → 0
great → +1

step 5: calculation :-
The movie was great !
+1+0 = 1
∴ since the polarity is greater than 0 so the givenstatement is '+'ve

textblob → python lib processing textual data & it allow to perform common NLP task such as POP tagging for, down phase extraction sentiment analysis classification

PACF → the plot summaises the correlation for an observation with lag values that is not accounted for prior lagged conditions

→ The Model is AR if the ACF trails off after a lag and has a hard cut off in the PACF after a lag
This lag is taken as the value for p.

→ The model is MA if the PACF " " " " " " " " " " ACF. This lag value taken as the ⇒ q

Step II → Estimation
involves using numerical methods to minimise a loss or error term
the method of least squared can be used

Step III ⇒ Diagonstic checking
→ Look for evidence that the model is not a good fit for the data
The two areas where DC is investigated are
(i) overfitting
(ii) Residual errors

what we do
↓
we start of checking if the model overfits the data
↓
The model is more complex than it needs to be & captures random noise in the training data
→ It negatively impacts the ability of the model to generalize, resulting in poor forecast performance on out of sample data

Forecast residuals provide a great opportun-ty for diagnostics
→ The error in model resemble white noise which is gaussian distribution with a me~~d~~an of zero & symmetrical variance
→ For this purpose use density plots, histogra[m] Q-Q plot that compare the distribution of error to the expected distribution.

Assumptions of ARIMA model

→ ~~stati~~ series is stationarity

→ Uncorrelated random error.

→ No outliers

→ Random shock (a random error component)

Steps to build Arimamodel

→

Box-Jenkins method → an iterative approach → 3 steps

Identification → Estimation → Diagonistic checking

step 1 :-

Access whether the
time series is stationary
and if not, how many
differences are required
to make it stationary

→

Identify the
parameter of an
Arima model for the
data

(a) use → unit root Tests → to determine whether or
not it is stationary

(b) Avoid over differencing.

steps of configuring AR and MA
2 digonstic plots can be choosen P, q. para
-meter for ~~AF~~ ~~to~~

→ Autocorrelation func (ACF)

→ Partial corre " (PACF)

→ Autocorrelation function (ACF) → The plot summarises the correlation of an observation with lag values. The x-axis shows the Lag and y-axis shows the correlation coeff b/w -1 & 1 for neg & pos correlatio