ORION
BUSINESS SOLUTIONS

Certificate of Approval

This is to certify the project report titled “ Analyzing Text and Generating Recommendations ” is a bonafide record of the work done by Dushyant Kumar Shukla at Orion Business Solutions, Indore, under the guidance of Mr. Ravindra Jain (Co-Founder and CEO) and senior developer Mr. Suddhatam Jain.

Orion Business Solutions

Mr. Ravindra Jain
(Co-Founder and CEO),
Orion Business Solutions

About Company

ORION
BUSINESS SOLUTIONS

Orin Business Solutions LLC. is an India-US based start-up which introduces PDAAS : ProductDevelopment-As-A-Service Platform alongside of SAAS, IAAS and PAAS . They will work with you as a team at each stage of product development to help you translate your ideas & strategies into marketable solutions.

At Orion Business Solutions, their endeavour is to help their clients grow by offering them innovative technological solutions, enhancing productivity and reducing timeto-market while maintaining optimal balance of cost, schedule and quality.

They are not typical outsourced IT company focused only on delivering codebase, which is just the tip of the iceberg. There are hidden parts of the product engineering crucial to a successful commercialization of an idea. They strive to help their clients on multiple stages of product development processes, starting from concept creation and all the way through post-production sustenance. They offer ProductDevelopment-as-a-Service with expertise in New Product Development (NPD) processes and focus on deep technological innovation.

They provide a unique and compelling alternative to typical team building and expanding options for start-ups and growing organizations, collaborating on each stage of Product Development Life Cycle. They distinguish themselves with deep technology competencies in software engineering services to allow you to focus on solving your business problems.

One of their primary differentiating factor is that most of them at Orion also have a business background and that coupled with their software engineering skills help them integrate business needs with technology in a seamless way.

Some of the areas they have expertise in (but not limited to) are Financial Trading Systems, E-Learning, E-Commerce, Social Media, Mobile Apps and UI/UX design. They have offices in Boston, MA and Indore, India.

Services provided at Orion Business Solutions

Custom Product Development

You must have heard of SAAS, IAAS and PAAS. Let us introduce you to PDAAS: Product Development-As-A-Service. We will work with you as a team at each stage of product development to help you translate your ideas & strategies into marketable solutions.

Venture Technology

Venture Capitalists invest money in a start-up. At Orion they invest their knowledge. They have over 50 years of combined experience in software development and that is our foundation for the pillar of your success.

Web Design

Uxmplify - That is what they call their User Interface design team. Their UI designers will make sure that your users love every pixel of your website. They will make sure the website renders content consistently irrespective of the viewing medium, be it a 55" screen or a 4" smartphone.

Mobile Apps

A software product is not complete unless you have a mobile interface for it and we excel at that. So far we have been focusing on developing native apps for Android and iOS only. Apart from that we also build apps using Phonegap/Cordova, and other cross-platform frameworks.

Cloud Deployment

Need to choose the right cloud solution? We can help strategize your cloud deployment, whether it is at Amazon's AWS, Microsoft Azure or other cloud hosting providers.

Big Data

Should we embrace Big Data? That's a dilemma a lot of the companies face these days. We can help you chalk out a Big Data strategy which is relevant to your business. We know Hadoop, NOSQL, and Recommendation Systems and can help you build a platform around these technologies.

Location Of the Company

Cambridge Innovation Centre
14th Floor, 1, Broadway, Cambridge, MA 02142 USA:
+1.919.937.9218

Orion Business Solutions
220, Trade Center, Kanchanbagh
South Tukoganj, Indore – 452001
India: +91.989.308.9895

info@orionbizsolutions.com

CHAPTER 1

RECOMMENDATION SYSTEM

Recommender systems or recommendation systems (sometimes replacing "system" with a synonym such as platform or engine) are a subclass of information filtering system that seek to predict the "rating" or "preference" that a user would give to an item.

Recommender systems have become extremely common in recent years, and are utilized in a variety of areas: some popular applications include movies, music, news, books, research articles, search queries, social tags, and products in general. There are also recommender systems for experts, collaborators, jokes, restaurants, garments, financial services, life insurance, romantic partners (online dating), and Twitter pages.

Algorithms that recommender systems use
As demonstrated by the winning approach for the Netflix prize, many algorithmic approaches are available for recommendation engines. Results can differ based on the problem the algorithm is designed to solve or the relationships that are present in the data. Many of the algorithms come from the field of machine learning, a subfield of artificial intelligence that produces algorithms for learning, prediction, and decision-making.

Pearson correlation
Similarity between two users (and their attributes, such as articles read from a collection of blogs) can be accurately calculated with the Pearson correlation. This algorithm measures the linear dependence between two variables (or users) as a function of their attributes. But it doesn't calculate this measure over the entire population of users. Instead, the population must be filtered down to neighbourhoods based on a higher-level similarity metric, such as reading similar blogs.
The Pearson correlation, which is widely used in research, is a popular algorithm for collaborative filtering.

1.1 Clustering algorithms

Clustering algorithms are a form of unsupervised learning that can find structure in a set of seemingly random (or unlabelled) data. In general, they work by identifying similarities among items, such as blog readers, by calculating their distance from other items in a feature space. (Features in a feature space could represent the number of articles read in a set of blogs.) The number of independent features defines the dimensionality of the space. If items are "close" together, they can be joined in a cluster.
Many clustering algorithms exist. The simplest one is k-means, which partitions items into k clusters. Initially, the items are randomly placed into clusters. Then, a centroid (or centre) is calculated for each cluster as a function of its members. Each item's distance from the centroids is then checked. If an item is found to be closer to another cluster, it's moved to that cluster. Centroids are recalculated each time all item distances are checked. When stability is reached (that is, when no items move during an iteration), the set is properly clustered, and the algorithm ends.
Calculating the distance between two objects can be difficult to visualize. One common method is to treat each item as a multidimensional vector and calculate the distance by using the Euclidean

algorithm. Other clustering variants include the Adaptive Resonance Theory (ART) family, Fuzzy Cmeans, and Expectation-Maximization (probabilistic clustering), to name a few.

Other algorithms
Many algorithms — and an even larger set of variations of those algorithms — exist for recommendation engines. Some that have been used successfully include:

- Bayesian Belief Nets, which can be visualized as a directed acyclic graph, with arcs representing the associated probabilities among the variables.

- Markov chains, which take a similar approach to Bayesian Belief Nets but treat the recommendation problem as sequential optimization instead of simply prediction.

- Rocchio classification (developed with the Vector Space Model), which exploits feedback of the item relevance to improve recommendation accuracy.

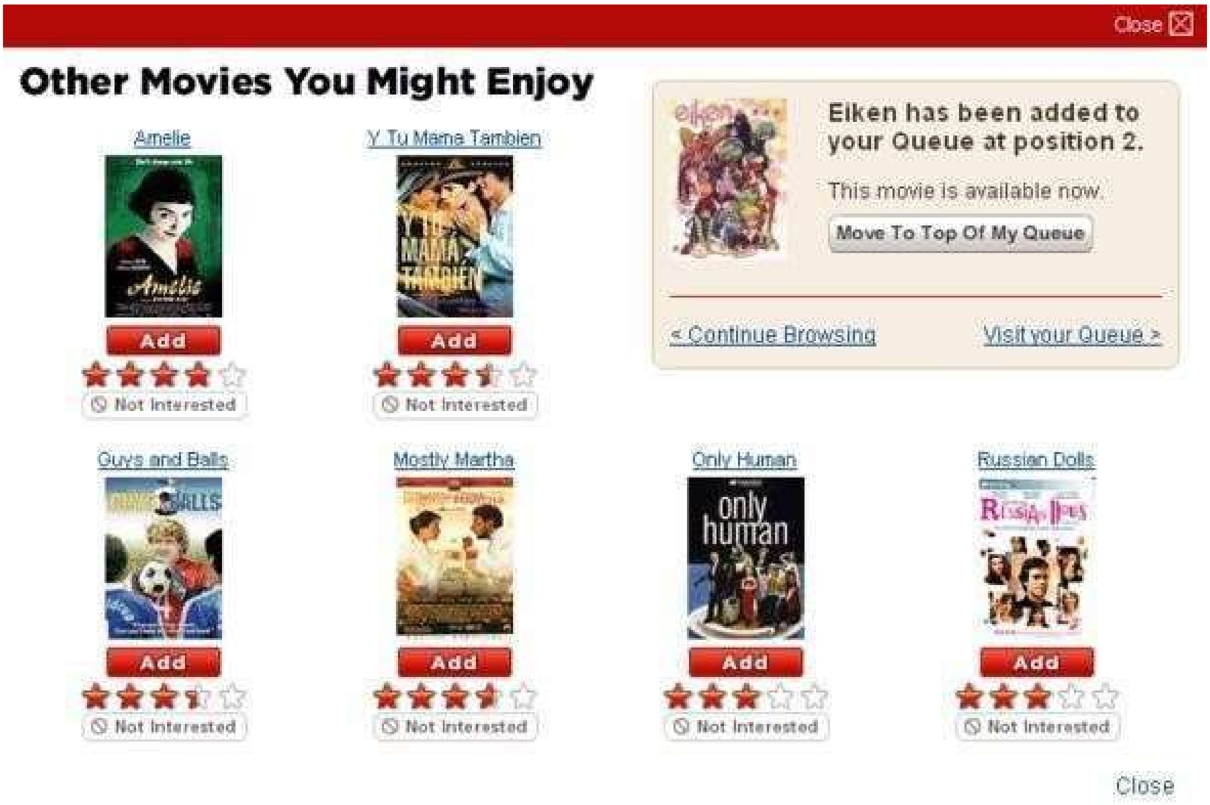
1.2 Challenges with recommender systems

Taking advantage of the "wisdom of crowds" (with collaborative filtering) has been made simpler with the data-collection opportunities the web affords. But the massive amounts of available data also complicate this opportunity. For example, although some users' behaviour can be modelled, other users do not exhibit typical behaviour. These users can skew the results of a recommender system and decrease its efficiency. Further, users can exploit a recommender system to favour one product over another — based on positive feedback on a product and negative feedback on competitive products, for example. A good recommender system must manage these issues.

One problem that's endemic to large-scale recommendation systems is scalability. Traditional algorithms work well with smaller amounts of data, but when the data sets grow, the traditional algorithms can have difficulty keeping up. Although this might not be a problem for offline processing, more-specialized approaches are needed for real-time scenarios.

Finally, privacy-protection considerations are also a challenge. Recommender algorithms can identify patterns individuals might not even know exist. A recent example is the case of a large company that could calculate a pregnancy-prediction score based on purchasing habits. Through the use of targeted ads, a father was surprised to learn that his teenage daughter was pregnant. The company's predictor was so accurate that it could predict a prospective mother's due date based on products she purchased.

NETFLIX:



Close ☒

Other Movies You Might Enjoy

Amelie

Add

Not Interested

Y Tu Mama Tambien

Add

Not Interested

Eiken has been added to your Queue at position 2.

This movie is available now.

Move To Top Of My Queue

< Continue Browsing

Visit your Queue >

Guys and Balls

Add

Not Interested

Mostly Martha

Add

Not Interested

Only Human

Add

Not Interested

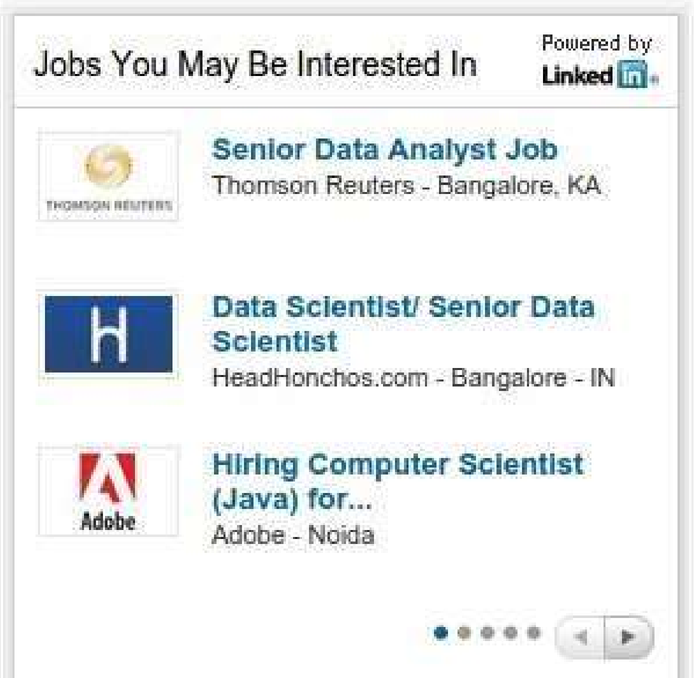
Russian Dolls

Add

Not Interested

Close

LINKEDIN:



Jobs You May Be Interested In

Powered by LinkedIn

Senior Data Analyst Job

Thomson Reuters - Bangalore, KA

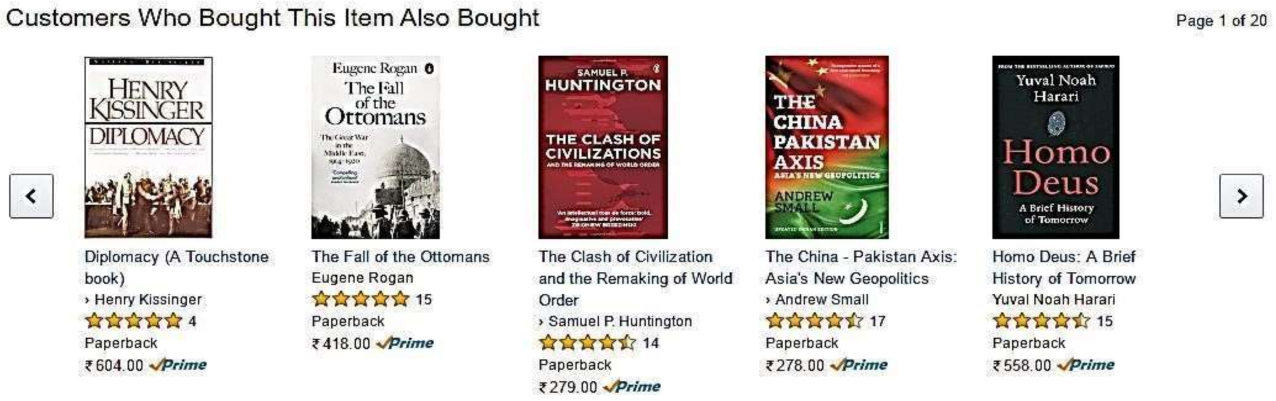
Data Scientist/ Senior Data Scientist

HeadHonchos.com - Bangalore - IN

Hiring Computer Scientist (Java) for...

Adobe - Noida

AMAZON:



Customers Who Bought This Item Also Bought

Page 1 of 20

Diplomacy (A Touchstone book)
› Henry Kissinger
★★★★★ 4
Paperback
₹604.00 ✓Prime

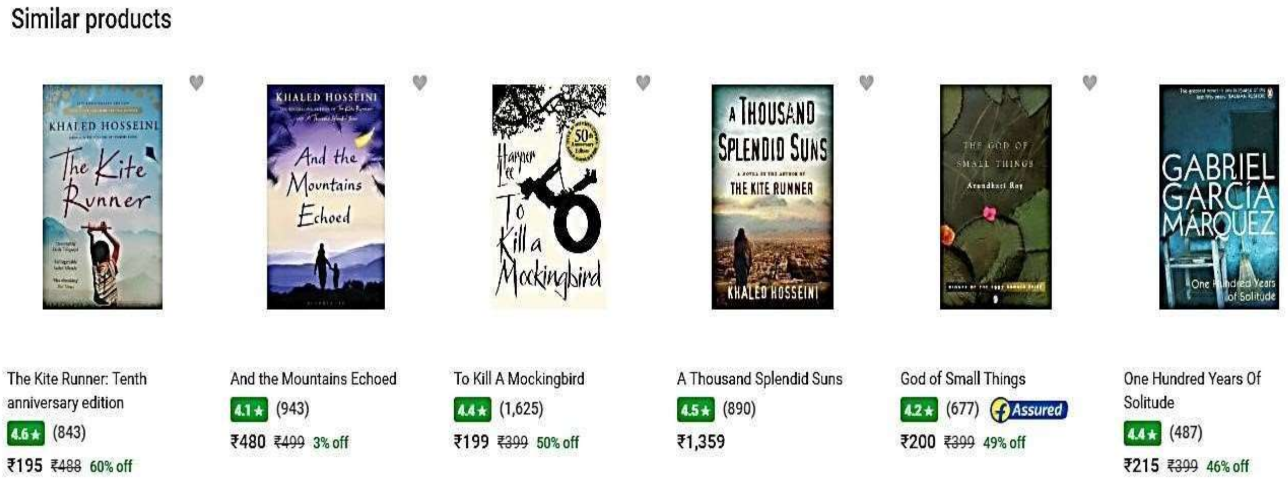
The Fall of the Ottomans
Eugene Rogan
★★★★★ 15
Paperback
₹418.00 ✓Prime

The Clash of Civilization and the Remaking of World Order
› Samuel P. Huntington
★★★★☆ 14
Paperback
₹279.00 ✓Prime

The China - Pakistan Axis: Asia's New Geopolitics
› Andrew Small
★★★★☆ 17
Paperback
₹278.00 ✓Prime

Homo Deus: A Brief History of Tomorrow
Yuval Noah Harari
★★★★☆ 15
Paperback
₹558.00 ✓Prime

FLIPKART:



Similar products

The Kite Runner: Tenth anniversary edition
4.6★ (843)
₹195 ~~₹488~~ 60% off

And the Mountains Echoed
4.1★ (943)
₹480 ~~₹499~~ 3% off

To Kill A Mockingbird
4.4★ (1,625)
₹199 ~~₹399~~ 50% off

A Thousand Splendid Suns
4.5★ (890)
₹1,359

God of Small Things
4.2★ (677) Assured
₹200 ~~₹399~~ 49% off

One Hundred Years Of Solitude
4.4★ (487)
₹215 ~~₹399~~ 46% off

CHAPTER 2
MACHINE LEARNING

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.

The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns. However, instead of extracting data for human comprehension -- as is the case in data mining applications -- machine learning uses that data to detect patterns in data and adjust program actions accordingly.

Machine learning algorithms are often categorized as being supervised or unsupervised. Supervised algorithms can apply what has been learned in the past to new data. Unsupervised algorithms can draw inferences from datasets.

Facebook's News Feed uses machine learning to personalize each member's feed. If a member frequently stops scrolling in order to read or "like" a particular friend's posts, the News Feed will start to show more of that friend's activity earlier in the feed. Behind the scenes, the software is simply using statistical analysis and predictive analytics to identify patterns in the user's data and use to patterns to populate the News Feed. Should the member no longer stop to read, like or comment on the friend's posts, that new data will be included in the data set and the News Feed will adjust accordingly.

2.1 Machine Learning: Algorithms Types

Machine learning algorithms are organized into taxonomy, based on the desired outcome of the algorithm. Common algorithm types include:

- Supervised learning --- where the algorithm generates a function that maps inputs to desired outputs. One standard formulation of the supervised learning task is the
- Classification problem: the learner is required to learn (to approximate the behaviour of) a function which maps a vector into one of several classes by looking at several input-output examples of the function.
- Unsupervised learning --- which models a set of inputs: labelled examples are not available.
- Semi-supervised learning --- which combines both labelled and unlabelled examples to generate an appropriate function or classifier.
- Reinforcement learning --- where the algorithm learns a policy of how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm.
- Transduction --- similar to supervised learning, but does not explicitly construct a function: instead, tries to predict new outputs based on training inputs, training outputs, and new inputs.
- Learning to learn --- where the algorithm learns its own inductive bias based on previous experience.

The performance and computational analysis of machine learning algorithms is a branch of statistics known as computational learning theory.

Machine learning is about designing algorithms that allow a computer to learn. Learning is not necessarily involves consciousness but learning is a matter of finding statistical regularities or other patterns in the data. Thus, many machine learning algorithms will barely resemble how human might approach a learning task. However, learning algorithms can give insight into the relative difficulty of learning in different environments.

2.2 Supervised Learning

Supervised learning is the machine learning task of inferring a function from labelled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyses the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

The parallel task in human and animal psychology is often referred to as concept learning.

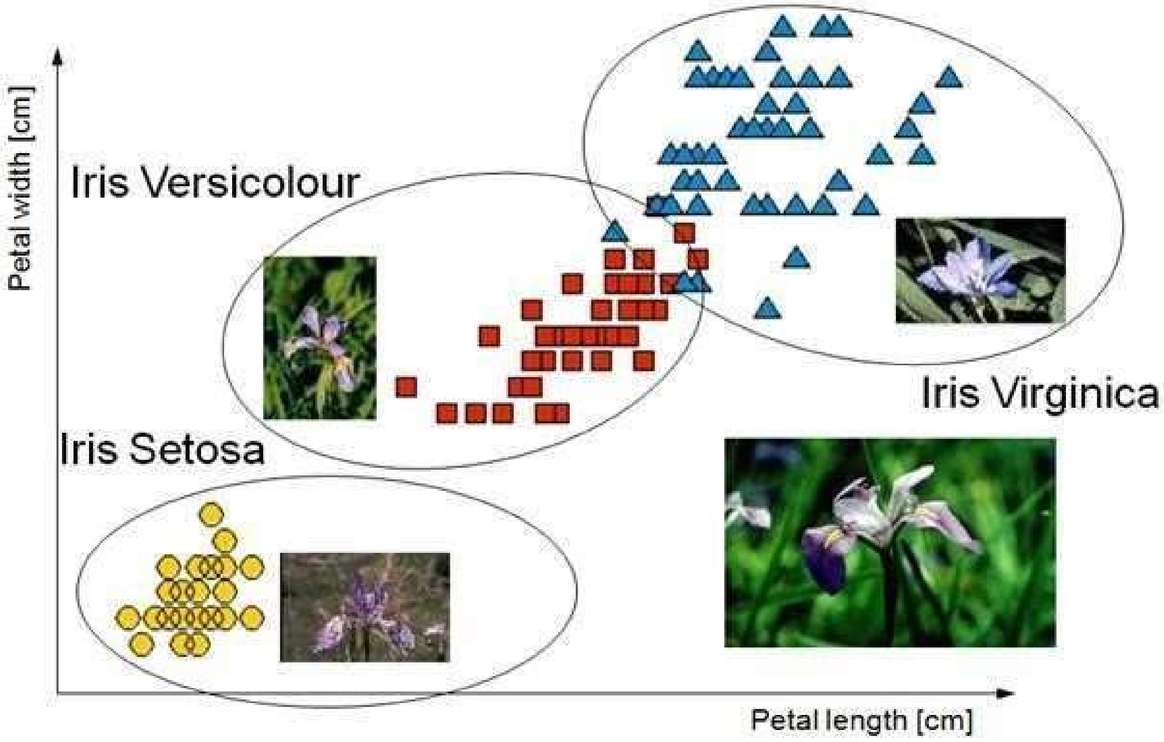
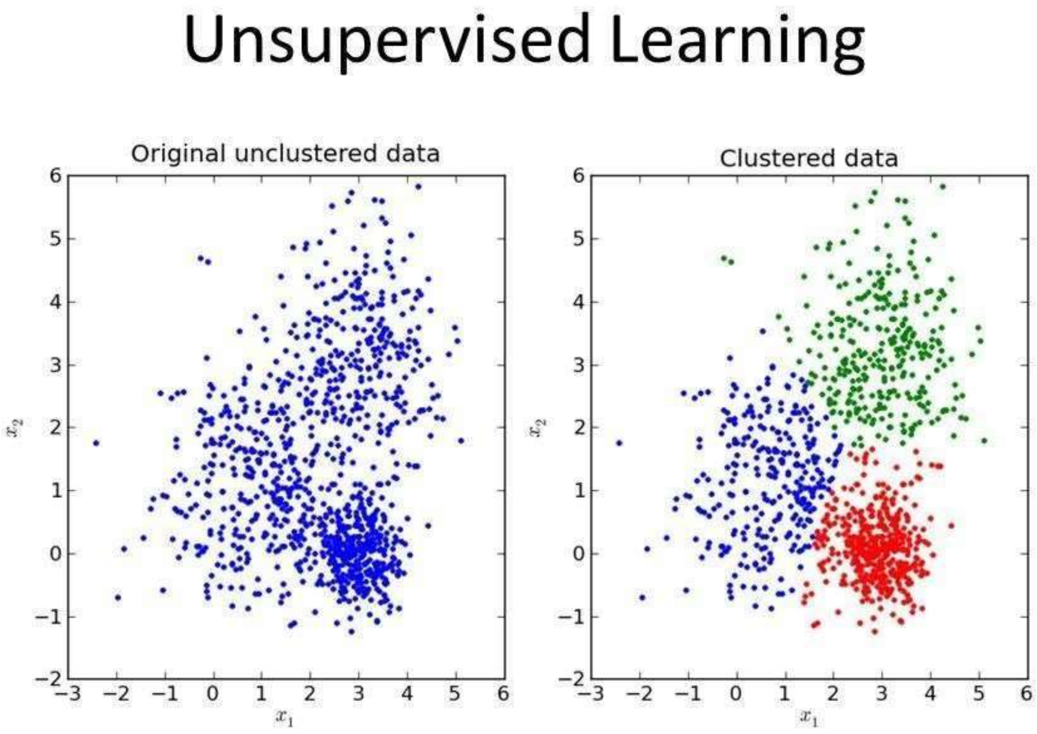


Fig. Visualization of Labelled Iris Dataset.

2.3 Unsupervised Learning

Unsupervised machine learning is the machine learning task of inferring a function to describe hidden structure from unlabelled data. Since the examples given to the learner are unlabelled, there is no error or reward signal to evaluate a potential solution – this distinguishes unsupervised learning from supervised learning and reinforcement learning.

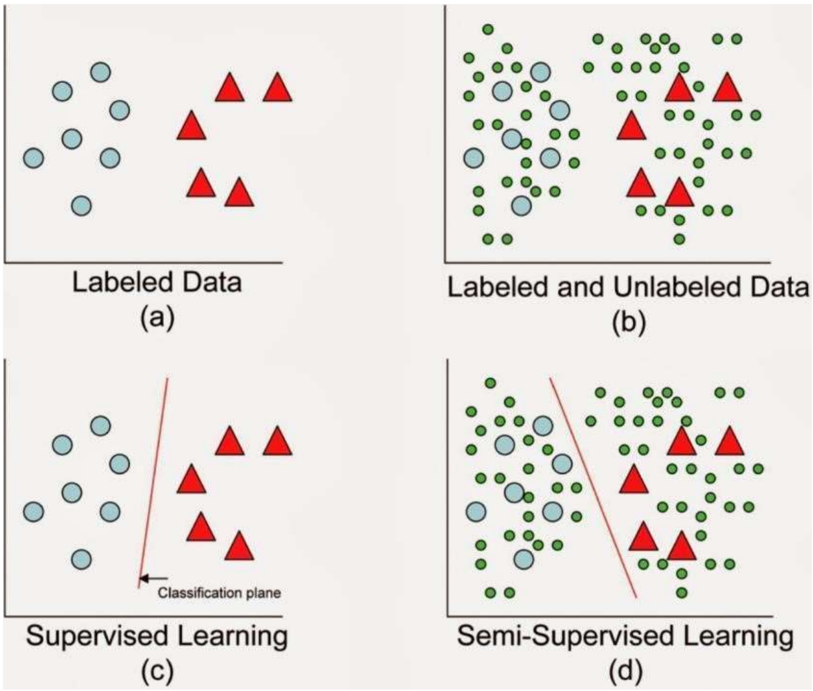
Unsupervised learning is closely related to the problem of density estimation in statistics. However, unsupervised learning also encompasses many other techniques that seek to summarize and explain key features of the data.



Unsupervised Learning

Fig. Representing both unlabelled data and clustered data

2.4 Semi-Supervised Learning



Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabelled data for training – typically a small amount of labelled data with a large amount of unlabelled data. Semi-supervised learning falls between unsupervised learning (without any labelled training data) and supervised learning (with completely labelled training data). Many machinelearning researchers have found that unlabelled data, when used in conjunction with a small amount of labelled data, can produce considerable improvement in learning accuracy. The acquisition of labelled data for a learning problem often requires a skilled human agent (e.g. to transcribe an audio segment) or a physical experiment (e.g. determining the 3D structure of a protein or determining whether there is oil at a particular location). The cost associated with the labelling process thus may render a fully labelled training set infeasible, whereas acquisition of unlabelled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value. Semisupervised learning is also of theoretical interest in machine learning and as a model for human learning.

CHAPTER 3
Text Analytics And Data Analytics Lifecycle

3.1 Introduction

Text analysis, sometimes called text analytics, refers to the representation, processing, and modelling of textual data to derive useful insights. An important component of text analysis is text mining, the process of discovering relationships and interesting patterns in large text collections.

Text analysis suffers from the curse of high dimensionality. Thus, we need different ways to reduce the dimensionality on random variables.

The process of reducing the number of random variables under consideration, via obtaining a set of principal variables is called dimensionality reduction or dimension reduction. It can be divided into feature selection and feature extraction.

We'll be using Feature Extraction method. Feature extraction transforms the data in the highdimensional space to a space of fewer dimensions.

3.2 Text Analysis Steps

A text analysis problem usually consists of three important steps: parsing, search and retrieval, and text mining.

Parsing is the process that takes unstructured text and imposes a structure for further analysis. The unstructured text could be a plain text file, a weblog, an Extensible Mark-up Language (XML) file, a Hypertext Mark-up Language (HTML) file, or a Word document. Parsing deconstructs the provided text and renders it in a more structured way for the subsequent steps.

Search and retrieval is the identification of the documents in a corpus that contain search items such as specific words, phrases, topics, or entities like people or organizations. These search items are generally called key terms. Search and retrieval originated from the field of library science and is now used extensively by web search engines.

Text-mining uses the terms and indexes produced by the prior two steps to discover meaningful insights pertaining to domains or problems of interest. With the proper representation of the text, many of the techniques, such as clustering and classification, can be adapted to text mining. For example, the k-means: Clustering, can be used to cluster text documents into groups, where each group represents a collect ion of documents with a similar topic. The distance of a document to a centroid represents how closely the document talks about that topic. Classification tasks such as sentiment analysis and spam filtering are prominent use cases for the naive Bayes classifier.

Text mining may utilize methods and techniques from various fields of study, such as statistical analysis, information retrieval, data mining, and natural language processing.
Note that, in reality, all three steps do not have to be present in a text analysis project. If the goal is to construct a corpus or provide a catalog service, for example, the focus would be the parsing task using one or more text pre-processing techniques, such as part-of-speech (POS) tagging, named entity

recognition, lemmatization, or stemming. Furthermore, the three tasks do not have to be sequential. Sometimes their orders might even look like a tree. For example, one could use parsing to build a data store and choose to either search and retrieve the related documents or use text mining on the entire data store to gain insights.

Since, we have been provided with structured database table of posts from our firm, we are only using Text-mining step on the entire data of posts table. We have also made use of Natural language processing for stemming the words.

3.3 Lemmatization, and Stemming

Both lemmatization and stemming are techniques to reduce the number of dimensions and reduce inflections or variant forms to the base form to more accurately measure the number of times each word appears.
With the use of a given dictionary, lemmatization finds the correct dictionary base form of a word.

For example, given the sentence: Obesity causes
many problems
The output of lemmatization would be:
Obesity cause many problem

Different from lemmatization, stemming does not need a dictionary, and it usually refers to a crude process of stripping affixes based on a set of heuristics with the hope of correctly achieving the goal to reduce inflections or variant forms. After the process, words are stripped to become stems. A stem is not necessarily an actual word defined in the natural language, but it is sufficient to differentiate itself from the stems of other words. A well-known rule-based stemming algorithm is Porter's stemming algorithm. It defines a set of production rules to iteratively transform words into their stems.

For the sentence shown previously:
obesity causes many problems the output of Porter's
stemming algorithm is:
obes caus mani problem

3.4 Data Analytics Lifecycle Diagram

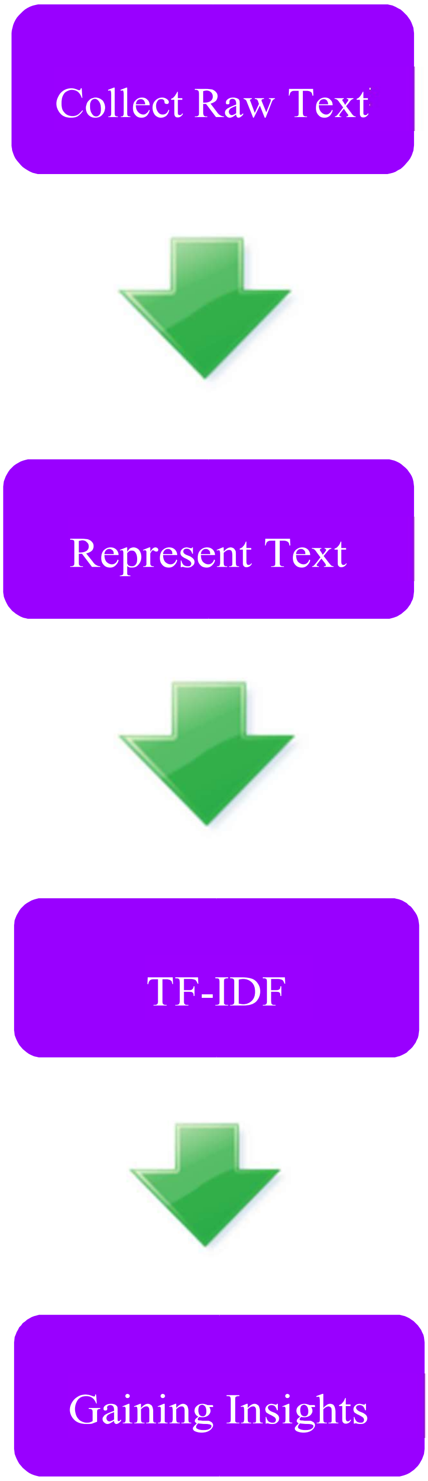


Figure showing Text Analysis Process which we have followed for our Project

Collect raw text

This corresponds to Phase 1 and Phase 2 of the Data Analytic Lifecycle. In this step, the Data Science team monitors websites for references to specific products. The websites may include social media and review sites. The team could interact with social network application programming interfaces (APIs) process data feeds, or scrape pages and use product names as keywords to get the raw data. Regular expressions are commonly used in this case to identify text that matches certain patterns. Additional filters can be applied to the raw data for a more focused study.

Represent text

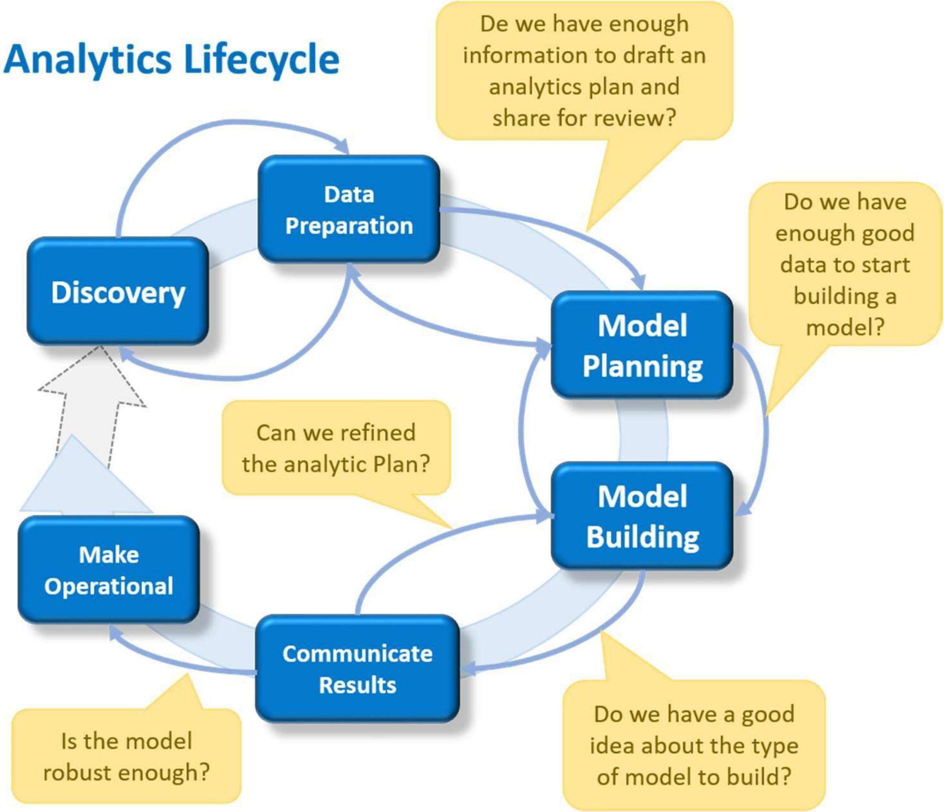
Convert each entry into a suitable document representation with proper indices, and build a corpus based on these indexed reviews. This step corresponds to Phases 2 and 3 of the Data Analytic Lifecycle.

TF-IDF

Compute the usefulness of each word in the entries using methods such as TFIDF. This step correspond to Phases 3 through 5 of the Data Analytic Lifecycle.

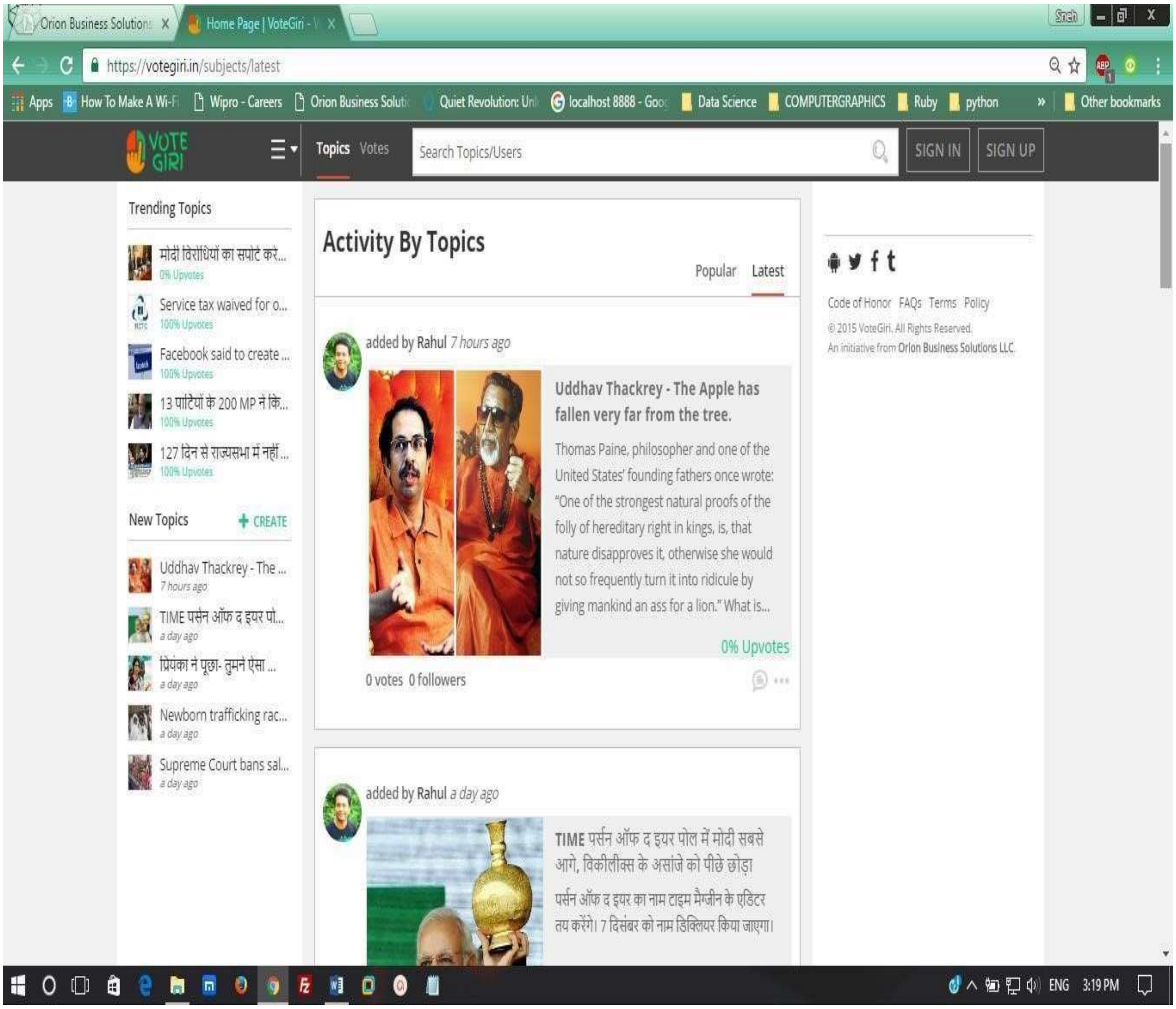
Gaining Insights

Review the results and gain greater insights. This step corresponds to Phase 5 and 6 of the Data Analytic Lifecycle. Test the soundness of the conclusions and operationalize the findings if applicable.


Analytics Lifecycle
De we have enough information to draft an analytics plan and share for review?
Data Preparation
Discovery
Do we have enough good data to start building a model?
Model Planning
Can we refined the analytic Plan?
Model Building
Make Operational
Communicate Results
Is the model robust enough?
Do we have a good idea about the type of model to build?

VOTEGIRI.IN

VoteGiri™ established in 2016 is a social voting platform for users to vote online and discuss on things that matter to them. It provides a way to measure the pulse of the nation by integrating a polling mechanism, and is a perfect way to see which way the country is swinging on particular issues. VoteGiri™ is a place for a conversation not just among friends and family, but also among like-minded individuals who are passionate about what is happening in the country.



Trending Topics

मोदी विरोधियों का सपोर्ट करे...
0% Upvotes

Service tax waived for o...
100% Upvotes

Facebook said to create ...
100% Upvotes

13 पार्टियों के 200 MP ने कि...
100% Upvotes

127 दिन से राज्यसभा में नहीं ...
100% Upvotes

New Topics

Uddhav Thackrey - The ...
7 hours ago

TIME पर्सन ऑफ द इयर पो...
a day ago

प्रियंका ने पूछा- तुमने ऐसा ...
a day ago

Newborn trafficking rac...
a day ago

Supreme Court bans sal...
a day ago

Activity By Topics

added by Rahul *7 hours ago*

Uddhav Thackrey - The Apple has fallen very far from the tree.

Thomas Paine, philosopher and one of the United States' founding fathers once wrote: "One of the strongest natural proofs of the folly of hereditary right in kings, is, that nature disapproves it, otherwise she would not so frequently turn it into ridicule by giving mankind an ass for a lion." What is...

0% Upvotes

0 votes 0 followers

added by Rahul *a day ago*

TIME पर्सन ऑफ द इयर पोल में मोदी सबसे आगे, विकीलीक्स के असांजे को पीछे छोड़ा

पर्सन ऑफ द इयर का नाम टाइम मैग्जीन के एडिटर तय करेंगे। 7 दिसंबर को नाम डिक्लियर किया जाएगा।

Code of Honor FAQs Terms Policy

© 2015 VoteGiri. All Rights Reserved.

An initiative from Orion Business Solutions LLC

CHAPTER 4
TOOLS AND LIBRARIES USED

- Python Language 2.7.12
- Pandas Library
- Jupyter Notebook
- Scipy Library
- scikit-learn Library
- nltk Library
- mysql.connector for python

4.1 Python Language and Pandas

Python has long been great for data munging and preparation, but less so for data analysis and modelling. Pandas helps fill this gap, enabling you to carry out your entire data analysis workflow in Python without having to switch to a more domain specific language like R.

Combined with the excellent Ipython toolkit and other libraries, the environment for doing data analysis in Python excels in performance, productivity, and the ability to collaborate.

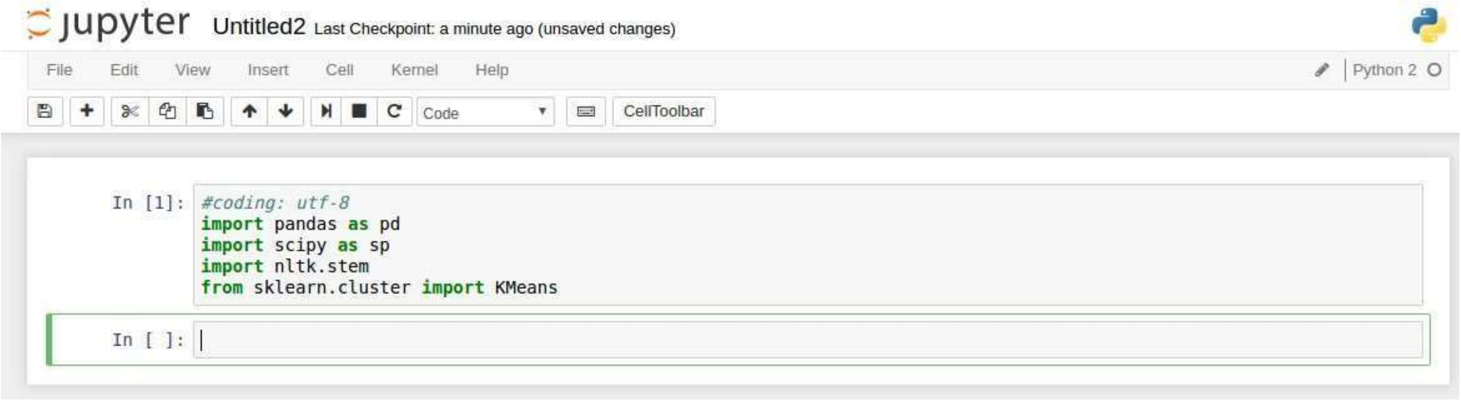
Pandas does not implement significant modelling functionality outside of linear and panel regression; for this, we have used scikit-learn .

Pandas Library Highlights

- A fast and efficient Data-Frame object for data manipulation with integrated indexing; Tools for reading and writing data between in-memory data structures and different formats: CSV and text files, Microsoft Excel, SQL databases, and the fast HDF5 format; Intelligent data alignment and integrated handling of missing data: gain automatic labelbased alignment in computations and easily manipulate messy data into an orderly form;
- Flexible reshaping and pivoting of data sets;
- Intelligent label-based slicing, fancy indexing, and sub-setting of large data sets; and much more.

4.2 The Jupyter Notebook

The Jupyter Notebook is a web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, machine learning and much more.



```
#coding: utf-8
import pandas as pd
import scipy as sp
import nltk.stem
from sklearn.cluster import KMeans
```

Language of choice

The Notebook has support for over 40 programming languages, including those popular in Data Science such as Python, R, Julia and Scala.

Big data integration

Leverage big data tools, such as Apache Spark, from Python, R and Scala. Explore that same data with pandas, scikit-learn, ggplot2, dplyr, etc.

Installation of Jupyter notebook

First, ensure that you have the latest pip; older versions may have trouble with some dependencies:

```
pip3 install --upgrade pip
```

Then install the Jupyter Notebook using:

```
pip3 install jupyter
```

4.3 Scipy Library

SciPy refers to several related but distinct entities:
The SciPy Stack, a collection of open source software for scientific computing in Python, and particularly a specified set of core packages.

□The community of people who use and develop this stack.

□Several conferences dedicated to scientific computing in Python - SciPy, EuroSciPy and SciPy.in.

□The SciPy library, one component of the SciPy stack, providing many numerical routines.

Core Packages

Python, a general purpose programming language. It is interpreted and dynamically typed and is very suited for interactive work and quick prototyping, while being powerful enough to write large applications in.

NumPy, the fundamental package for numerical computation. It defines the numerical array and matrix types and basic operations on them.

The SciPy library, a collection of numerical algorithms and domain-specific toolboxes, including signal processing, optimization, statistics and much more.

Matplotlib, a mature and popular plotting package, that provides publication-quality 2D plotting as well as rudimentary 3D plotting

pandas, providing high-performance, easy to use data structures.

SymPy, for symbolic mathematics and computer algebra.

4.4 scikit-learn

Machine Learning in Python

Simple and efficient tools for data mining and data analysis

Accessible to everybody, and reusable in various contexts

Built on NumPy, SciPy, and matplotlib

Open source, commercially usable - BSD license

We have made use of this mainly because of its feature like:

Clustering: Automatic grouping of similar objects into sets.

- Applications: Customer segmentation, Grouping experiment outcomes Algorithms: kMeans, spectral clustering, mean-shift etc.

Pre-processing: Feature extraction and normalization.

- Application: Transforming input data such as text for use with machine learning algorithms.
- Modules: pre-processing, feature extraction.

Dimensionality Reduction: Reducing the number of random variables to consider.

- Applications: Visualization, Increased efficiency
- Algorithms: PCA, feature selection, non-negative matrix factorization.

4.5 Natural Language Toolkit

We have made use of this for the sole purpose of Stemming the words.

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analysing linguistic structure, and more.

Installing NLTK
NLTK requires Python versions 2.7 or 3.2+

Mac/Unix

1.Install NLTK: run sudo pip install -U nltk
2.Install Numpy (optional): run sudo pip install -U numpy
3.Test installation: run python then type import nltk

4.6 MySQL Connector/Python

MySQL Connector/Python enables Python programs to access MySQL databases, using an API that is compliant with the Python Database API Specification v2.0 (PEP 249). It is written in pure Python and does not have any dependencies except for the Python Standard Library.

MySQL Connector/Python includes support for:

- Almost all features provided by MySQL Server up to and including MySQL Server version 5.7.
- Converting parameter values back and forth between Python and MySQL data types, for example Python date time and MySQL DATETIME. You can turn automatic conversion on for convenience, or off for optimal performance.
- All MySQL extensions to standard SQL syntax.
- Protocol compression, which enables compressing the data stream between the client and server.
- Connections using TCP/IP sockets and on UNIX using UNIX sockets.
- Secure TCP/IP connections using SSL.
- Self-contained driver. Connector/Python does not require the MySQL client library or any Python modules outside the standard library.

CHAPTER 5
Building Recommendation system for Votegiri.in

To generate recommendation first we need to find some patterns in the Posts. When a user browses our site looking for some particular information, the search engine will most likely point him/her to a specific post. To improve the user experience, we wanted to show all related post with their texts. If the presented post is not what he/she was looking for, he/she can easily see the other available posts and hopefully stay on our site.

The naive approach would be to take the post, calculate its similarity to all other posts, and display the top N most similar posts as links on the page. This will quickly become very costly. Instead, we need a method that quickly finds all related posts.

We achieved this goal by using clustering. This is a method of arranging items so that similar items are in one cluster and dissimilar items are in distinct ones.

The tricky thing that we have to tackle first is how to turn text into something on which we can calculate similarity. With such a measurement for similarity, we will then proceed to investigate how we can leverage that to quickly arrive at a cluster that contains similar posts.
Once there, we will only have to check out those documents that also belong to that cluster. To achieve this, we will use the marvellous Scikit library, which comes with diverse machine-learning methods.

We will use the Bag of Words approach. Bag-of-words takes quite a naive approach, as order plays an important role in the semantics of text. With bag-of-words, many texts with different meanings are combined into one form. For example, the texts "a dog bites a man" and "a man bites a dog" have very different meanings, but they would share the same representation with bag-of-words. Although the bag-of-words technique oversimplifies the problem, it is still considered a good approach to start with, and it is widely used for text analysis.

It uses simple word counts or word frequencies as its basis. For each word in the post, its occurrence is counted and noted in a vector. Not surprisingly, this step is also called vectorization. The vector is typically huge as it contains as many elements as the words that occur in the whole dataset.

Take for instance two example posts with the following word
counts: Word   Occurrences in Post 1   Occurrences in Post 2 disk
1   1 format   1   1 how   1
0 hard   1   1 my   1   0 problems
0   1 to   1   0
The columns Post 1 and Post 2 can now be treated as simple vectors. We could simply calculate the Euclidean distance between the vectors of all posts and take the nearest one (too slow, as we have just found out). As such, we can use them later in the form of feature vectors in the following clustering steps:

1. Extract the salient features from each post and store it as a vector per post.
2. Compute clustering on the vectors.
3. Determine the cluster for the post in question.
4. From this cluster, fetch a handful of posts that are different from the post in question. This will increase diversity.

However, there is some more work to be done before we get there, and before we can do that work, we need some data to work on. For the purpose of vectorization of words present in each post we have made use of the TF-IDF vectorizer from scikit-learn library. The tf-idf vectorizer convert a collection of raw documents to a matrix of TF-IDF features. Equivalent to CountVectorizer followed by TfidfTransformer.

5.1 Calculating tf-idf (step 1)

In information retrieval, tf–idf, short for term frequency–inverse document frequency.

It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

Variations of the tf–idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. Tf–idf can be successfully used for stopwords filtering in various subject fields including text summarization and classification.

One of the simplest ranking functions is computed by summing the tf–idf for each query term; many more sophisticated ranking functions are variants of this simple model.

Term frequency

Suppose we have a set of English text documents and wish to determine which document is most relevant to the query "the brown cow". A simple way to start out is by eliminating documents that do not contain all three words "the", "brown", and "cow", but this still leaves many documents.

To further distinguish them, we might count the number of times each term occurs in each document and sum them all together; the number of times a term occurs in a document is called its term frequency.

Inverse document frequency

Because the term "the" is so common, term frequency will tend to incorrectly emphasize documents which happen to use the word "the" more frequently, without giving enough weight to the more meaningful terms "brown" and "cow".

The term "the" is not a good keyword to distinguish relevant and non-relevant documents and terms, unlike the less common words "brown" and "cow".

Hence, an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

How to Compute:

Typically, the tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

- TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

- IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

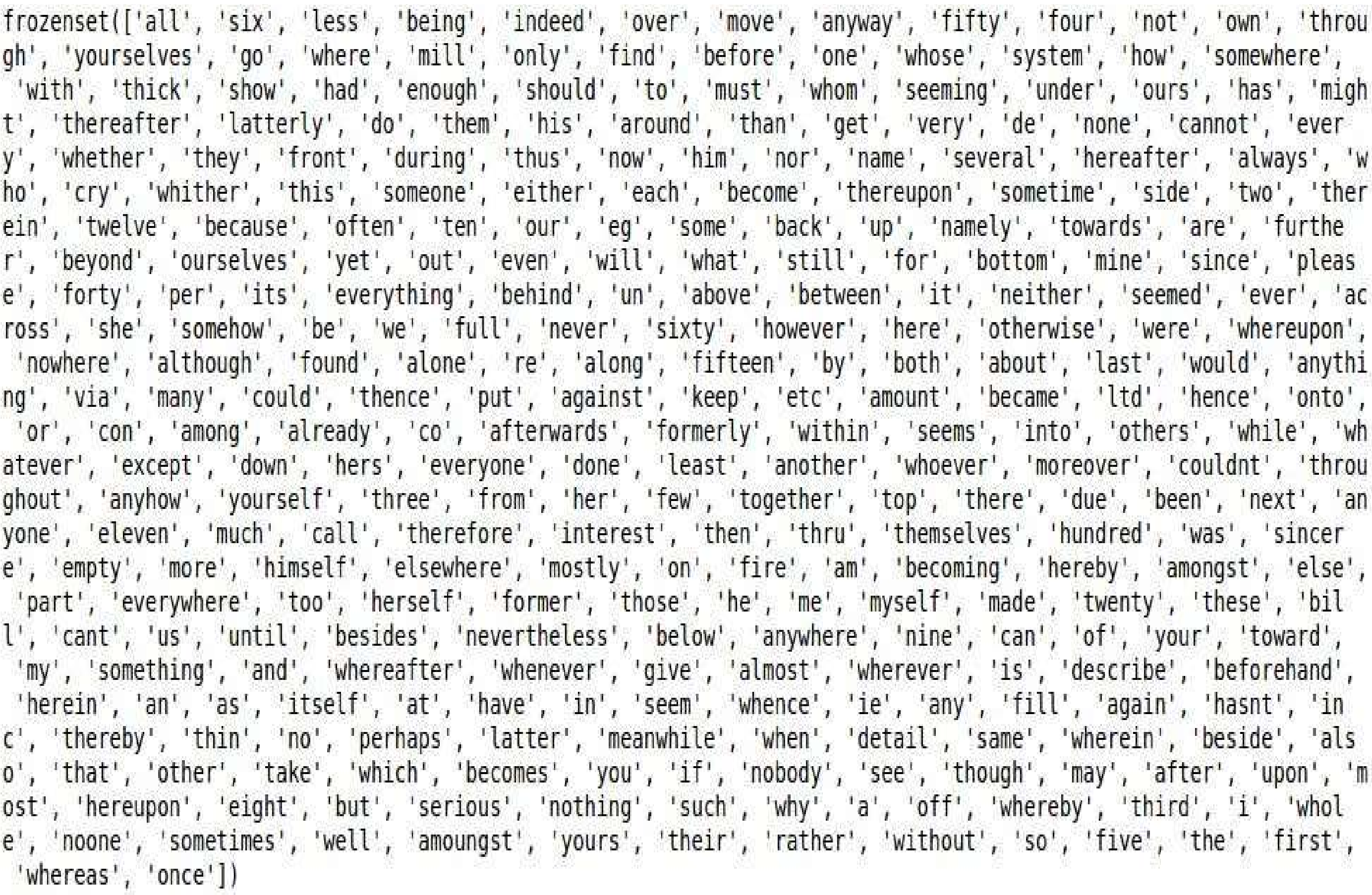
  IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

See below for a simple example.

Example:

Consider a document containing 100 words wherein the word cat appears 3 times. The term frequency (i.e., tf) for cat is then (3 / 100) = 0.03. Now, assume we have 10 million documents and the word cat appears in one thousand of these. Then, the inverse document frequency (i.e., idf) is calculated as log(10,000,000 / 1,000) = 4. Thus, the Tf-idf weight is the product of these quantities:
0.03 * 4 = 0.12.

5.2 Removing less important words (step 2)



```
frozenset(['all', 'six', 'less', 'being', 'indeed', 'over', 'move', 'anyway', 'fifty', 'four', 'not', 'own', 'throu
gh', 'yourselves', 'go', 'where', 'mill', 'only', 'find', 'before', 'one', 'whose', 'system', 'how', 'somewhere',
 'with', 'thick', 'show', 'had', 'enough', 'should', 'to', 'must', 'whom', 'seeming', 'under', 'ours', 'has', 'migh
t', 'thereafter', 'latterly', 'do', 'them', 'his', 'around', 'than', 'get', 'very', 'de', 'none', 'cannot', 'ever
y', 'whether', 'they', 'front', 'during', 'thus', 'now', 'him', 'nor', 'name', 'several', 'hereafter', 'always', 'w
ho', 'cry', 'whither', 'this', 'someone', 'either', 'each', 'become', 'thereupon', 'sometime', 'side', 'two', 'ther
ein', 'twelve', 'because', 'often', 'ten', 'our', 'eg', 'some', 'back', 'up', 'namely', 'towards', 'are', 'furthe
r', 'beyond', 'ourselves', 'yet', 'out', 'even', 'will', 'what', 'still', 'for', 'bottom', 'mine', 'since', 'pleas
e', 'forty', 'per', 'its', 'everything', 'behind', 'un', 'above', 'between', 'it', 'neither', 'seemed', 'ever', 'ac
ross', 'she', 'somehow', 'be', 'we', 'full', 'never', 'sixty', 'however', 'here', 'otherwise', 'were', 'whereupon',
 'nowhere', 'although', 'found', 'alone', 're', 'along', 'fifteen', 'by', 'both', 'about', 'last', 'would', 'anythi
ng', 'via', 'many', 'could', 'thence', 'put', 'against', 'keep', 'etc', 'amount', 'became', 'ltd', 'hence', 'onto',
 'or', 'con', 'among', 'already', 'co', 'afterwards', 'formerly', 'within', 'seems', 'into', 'others', 'while', 'wh
atever', 'except', 'down', 'hers', 'everyone', 'done', 'least', 'another', 'whoever', 'moreover', 'couldnt', 'throu
ghout', 'anyhow', 'yourself', 'three', 'from', 'her', 'few', 'together', 'top', 'there', 'due', 'been', 'next', 'an
yone', 'eleven', 'much', 'call', 'therefore', 'interest', 'then', 'thru', 'themselves', 'hundred', 'was', 'sincer
e', 'empty', 'more', 'himself', 'elsewhere', 'mostly', 'on', 'fire', 'am', 'becoming', 'hereby', 'amongst', 'else',
 'part', 'everywhere', 'too', 'herself', 'former', 'those', 'he', 'me', 'myself', 'made', 'twenty', 'these', 'bil
l', 'cant', 'us', 'until', 'besides', 'nevertheless', 'below', 'anywhere', 'nine', 'can', 'of', 'your', 'toward',
 'my', 'something', 'and', 'whereafter', 'whenever', 'give', 'almost', 'wherever', 'is', 'describe', 'beforehand',
 'herein', 'an', 'as', 'itself', 'at', 'have', 'in', 'seem', 'whence', 'ie', 'any', 'fill', 'again', 'hasnt', 'in
c', 'thereby', 'thin', 'no', 'perhaps', 'latter', 'meanwhile', 'when', 'detail', 'same', 'wherein', 'beside', 'als
o', 'that', 'other', 'take', 'which', 'becomes', 'you', 'if', 'nobody', 'see', 'though', 'may', 'after', 'upon', 'm
ost', 'hereupon', 'eight', 'but', 'serious', 'nothing', 'such', 'why', 'a', 'off', 'whereby', 'third', 'i', 'whol
e', 'noone', 'sometimes', 'well', 'amoungst', 'yours', 'their', 'rather', 'without', 'so', 'five', 'the', 'first',
 'whereas', 'once'])
```

Fig. Representing list of stop words

```
>>> vectorizer = CountVectorizer(stop_words='english')
```

Let us have another look at Post. Of its words that are not in the new post, we have "most", "safe", "images", and "permanently". They are actually quite different in the overall importance to the post.

Words such as "most" appear very often in all sorts of different contexts, and words such as this are called stop words. They do not carry as much information, and thus should not be weighed as much as words such as "images", that don't occur often in different contexts.

The best option would be to remove all words that are so frequent that they do not help to distinguish between different texts. These words are called stop words.

As this is such a common step in text processing, there is a simple parameter in tfidfVectorizer to achieve this.

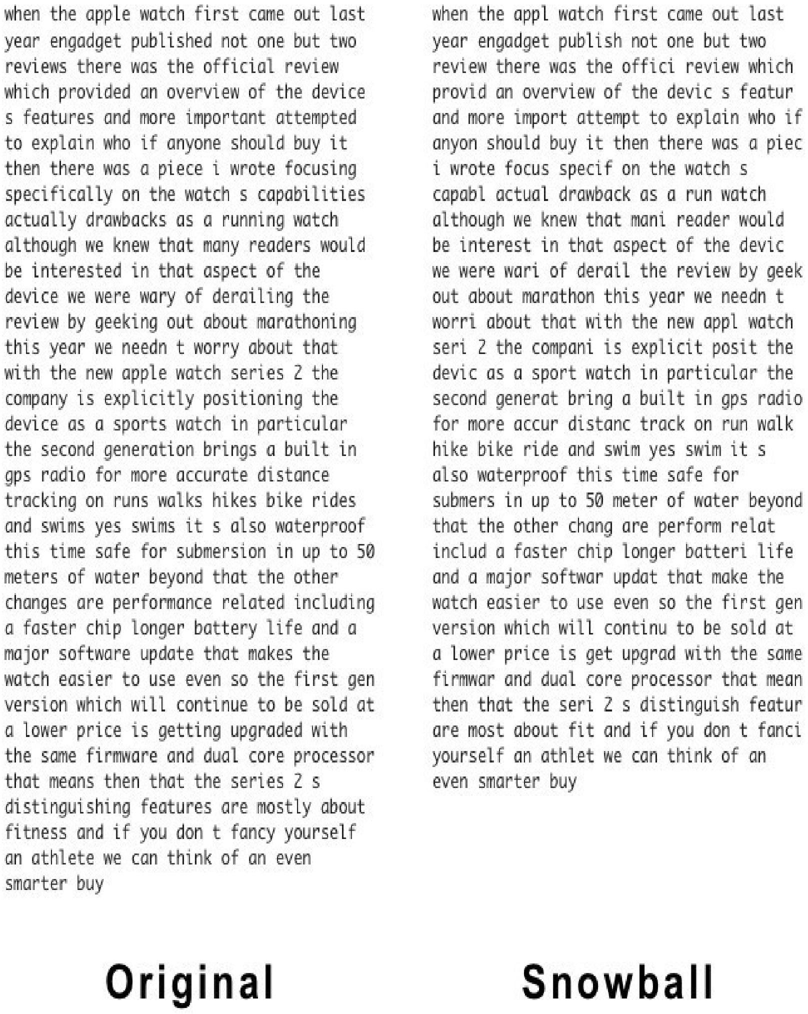
5.3 Stemming (step 3)

One thing is still missing. We count similar words in different variants as different words. For instance, a post contains "imaging" and "images" words. It would make sense to count them together. After all, it is the same concept they are referring to.

We need a function that reduces words to their specific word stem. Scikit does not contain a stemmer by default. With the Natural Language Toolkit (NLTK), we can download a free software toolkit, which provides a stemmer that we can easily plug into our tfidfVectorizer.

NLTK comes with different stemmers. This is necessary, because every language has a different set of rules for stemming. For English, we can take Snowball Stemmer.

```
>>> import nltk.stem
>>> s= nltk.stem.SnowballStemmer('english')
>>> s.stem("graphics") u'graphic'
```



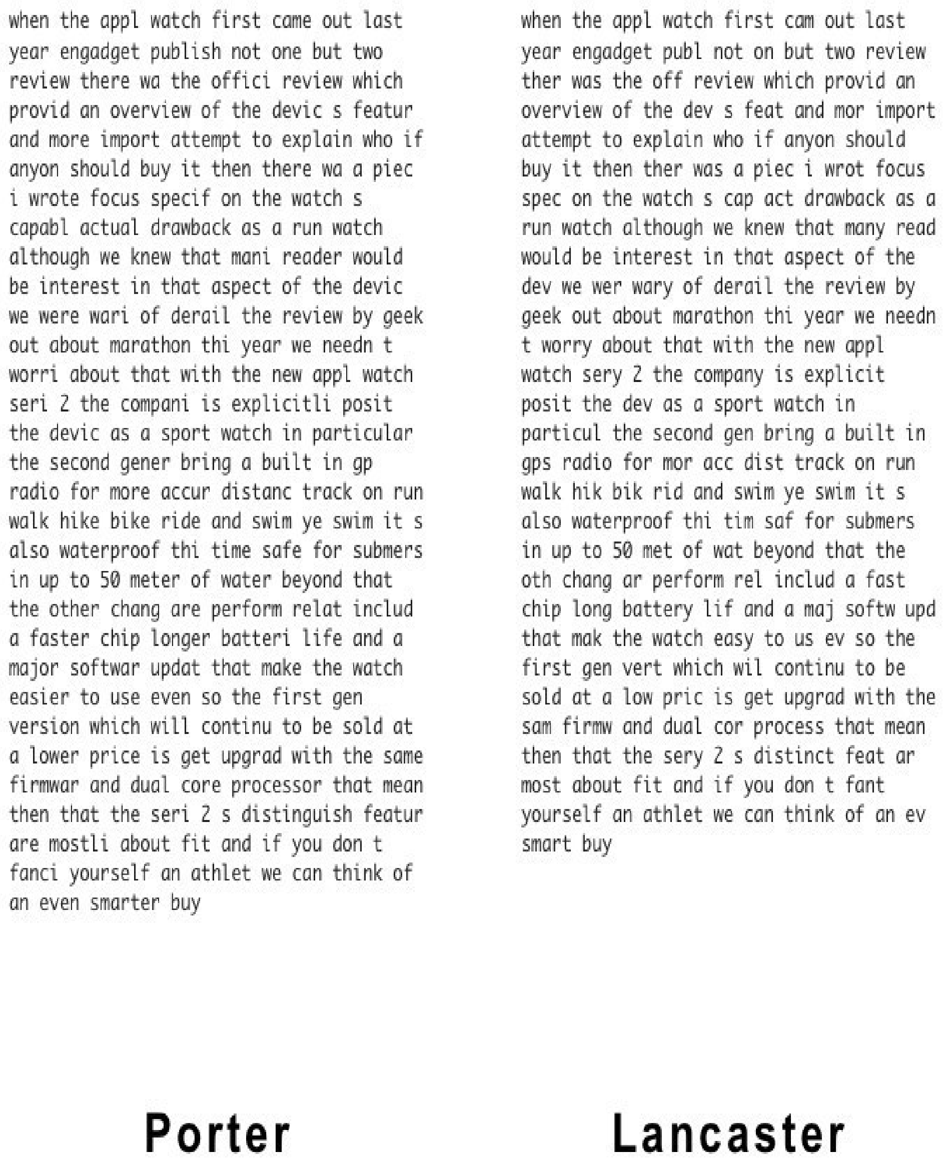
when the apple watch first came out last year engadget published not one but two reviews there was the official review which provided an overview of the device s features and more important attempted to explain who if anyone should buy it then there was a piece i wrote focusing specifically on the watch s capabilities actually drawbacks as a running watch although we knew that many readers would be interested in that aspect of the device we were wary of derailing the review by geeking out about marathoning this year we needn t worry about that with the new apple watch series 2 the company is explicitly positioning the device as a sports watch in particular the second generation brings a built in gps radio for more accurate distance tracking on runs walks hikes bike rides and swims yes swims it s also waterproof this time safe for submersion in up to 50 meters of water beyond that the other changes are performance related including a faster chip longer battery life and a major software update that makes the watch easier to use even so the first gen version which will continue to be sold at a lower price is getting upgraded with the same firmware and dual core processor that means then that the series 2 s distinguishing features are mostly about fitness and if you don t fancy yourself an athlete we can think of an even smarter buy

when the appl watch first came out last year engadget publish not one but two review there was the offici review which provid an overview of the devic s featur and more import attempt to explain who if anyon should buy it then there was a piec i wrote focus specif on the watch s capabl actual drawback as a run watch although we knew that mani reader would be interest in that aspect of the devic we were wari of derail the review by geek out about marathon this year we needn t worri about that with the new appl watch seri 2 the compani is explicit posit the devic as a sport watch in particular the second generat bring a built in gps radio for more accur distanc track on run walk hike bike ride and swim yes swim it s also waterproof this time safe for submers in up to 50 meter of water beyond that the other chang are perform relat includ a faster chip longer batteri life and a major softwar updat that make the watch easier to use even so the first gen version which will continu to be sold at a lower price is get upgrad with the same firmwar and dual core processor that mean then that the seri 2 s distinguish featur are most about fit and if you don t fanci yourself an athlet we can think of an even smarter buy

Original

Snowball

Fig. When using Snowball stemmer



Porter

```
when the appl watch first came out last
year engadget publish not one but two
review there wa the offici review which
provid an overview of the devic s featur
and more import attempt to explain who if
anyon should buy it then there wa a piec
i wrote focus specif on the watch s
capabl actual drawback as a run watch
although we knew that mani reader would
be interest in that aspect of the devic
we were wari of derail the review by geek
out about marathon thi year we needn t
worri about that with the new appl watch
seri 2 the compani is explicitli posit
the devic as a sport watch in particular
the second gener bring a built in gp
radio for more accur distanc track on run
walk hike bike ride and swim ye swim it s
also waterproof thi time safe for submers
in up to 50 meter of water beyond that
the other chang are perform relat includ
a faster chip longer batteri life and a
major softwar updat that make the watch
easier to use even so the first gen
version which will continu to be sold at
a lower price is get upgrad with the same
firmwar and dual core processor that mean
then that the seri 2 s distinguish featur
are mostli about fit and if you don t
fanci yourself an athlet we can think of
an even smarter buy
```

Lancaster

```
when the appl watch first cam out last
year engadget publ not on but two review
ther was the off review which provid an
overview of the dev s feat and mor import
attempt to explain who if anyon should
buy it then ther was a piec i wrot focus
spec on the watch s cap act drawback as a
run watch although we knew that many read
would be interest in that aspect of the
dev we wer wary of derail the review by
geek out about marathon thi year we needn
t worry about that with the new appl
watch sery 2 the company is explicit
posit the dev as a sport watch in
particul the second gen bring a built in
gps radio for mor acc dist track on run
walk hik bik rid and swim ye swim it s
also waterproof thi tim saf for submers
in up to 50 met of wat beyond that the
oth chang ar perform rel includ a fast
chip long battery lif and a maj softw upd
that mak the watch easy to us ev so the
first gen vert which wil continu to be
sold at a low pric is get upgrad with the
sam firmw and dual cor process that mean
then that the sery 2 s distinct feat ar
most about fit and if you don t fant
yourself an athlet we can think of an ev
smart buy
```

Fig. When using Porter or Lancaster Stemmers

What we have achieved till now?

Our current text pre-processing phase includes the following steps:

1. Tokenizing the text.

2. Throwing away words that occur way too often to be of any help in detecting relevant posts.

3. Throwing away words that occur so seldom that there is only a small chance that they occur in future posts.

4. Counting the remaining words.

5. Calculating TF-IDF values from the counts, considering the whole text corpus.

In simple words, we have converted all our posts to vectors, which are represented by a vectorise array, and each element of that array is vector-matrix.
With this process, we are able to convert a bunch of noisy text into a concise representation of feature values.

5.4 Clustering (Unsupervised Machine Learning) (step 4)

Overview

In general, clustering is the use of unsupervised techniques for grouping similar objects. In machine learning, unsupervised refers to the problem of finding hidden structure within unlabelled data. Clustering techniques are unsupervised in the sense that the data scientist does not determine, in advance, the labels to apply to the clusters. The structure of the data describes the objects of interest and determines how best to group the objects. For example, based on customers' personal income, it is straightforward to divide the customers into three groups depending on arbitrarily selected values. The customers could be divided into three groups as follows:

- Earn less than $10,000
- Earn between 510,000 and $99,999
- Earn $100,000 or more

In this case, the income levels were chosen somewhat subjectively based on easy-to-communicate points of delineation. However, such groupings do not indicate a natural affinity of the customers within each group. In other words, there is no inherent reason to believe that the customer making $90,000 will behave any differently than the customer making 5110,000. As additional dimensions are introduced by adding more variables about the customers, the task of finding meaningful groupings becomes more complex. For instance, suppose variables such as age, years of education, household size, and annual purchase expenditures were considered along with the personal income variable. What are the natural occurring groupings of customers? This is the type of question that clustering analysis can help answer.

Clustering is a method often used for exploratory ana lysis of the data. In clustering, there are no predictions made. Rather, clustering methods find the similarities between objects according to the object attributes and group the similar objects into clusters. Clustering techniques are utilized in marketing, economics, and various branches of science. A popular clustering method is k-means. Since, we have our vectors that we believe capture the posts to a sufficient degree.

Not surprisingly, there are many ways to group them together. Most clustering algorithms fall into one of the two methods, flat and hierarchical clustering.

Flat clustering divides the posts into a set of clusters without relating the clusters to each other. The goal is simply to come up with a partitioning such that all posts in one cluster are most similar to each other while being dissimilar from the posts in all other clusters. Many flat clustering algorithms require the number of clusters to be specified up front.

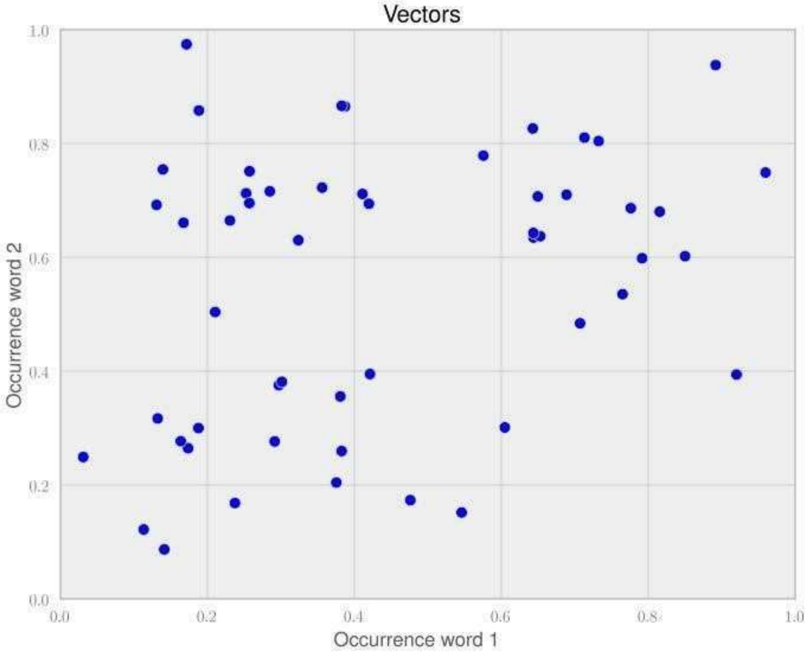
In hierarchical clustering, the number of clusters does not have to be specified. Instead, the hierarchical clustering creates a hierarchy of clusters. While similar posts are grouped into one cluster, similar clusters are again grouped into one uber-cluster. This is done recursively, until only one cluster is left, which contains everything. In this hierarchy, one can then choose the desired number of clusters. However, this comes at the cost of lower efficiency.

Scikit provides a wide range of clustering approaches in the package sklearn.cluster. You can get a quick overview of the advantages and drawbacks of each of them at http://scikitlearn.org/dev/modules/clustering.html.
In the following section, we will use the flat clustering method, KMeans, and play a bit with the desired number of clusters.

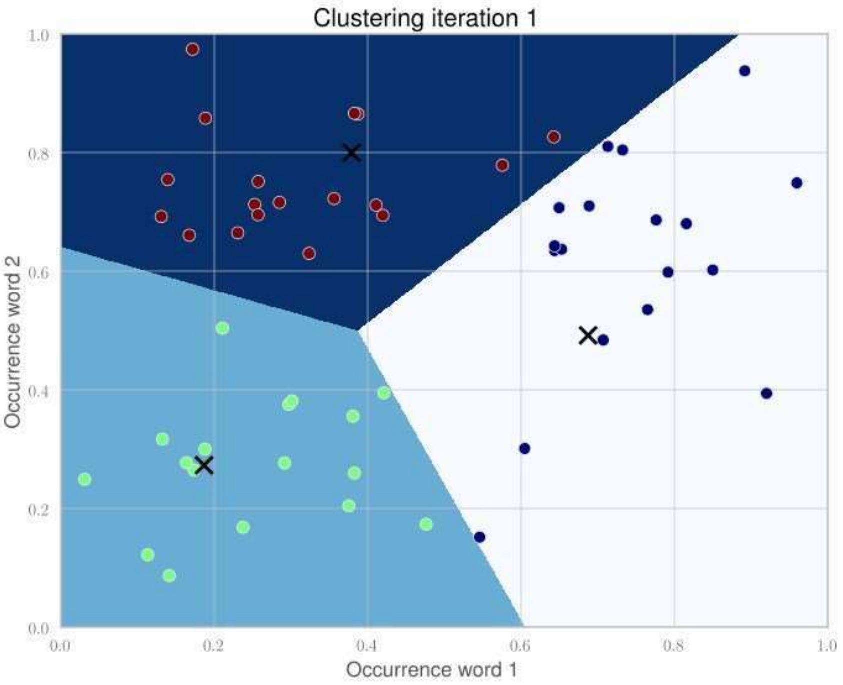
5.4.1 KMeans

KMeans is the most widely used flat clustering algorithm. After it is initialized with the desired number of clusters, num_clusters , it maintains that number of so-called cluster centroids. Initially, it would pick any of the num_clusters posts and set the centroids to their feature vector. Then it would go through all other posts and assign them the nearest centroid as their current cluster. Then it will move each centroid into the middle of all the vectors of that particular class. This changes, of course, the cluster assignment. Some posts are now nearer to another cluster. So it will update the assignments for those changed posts. This is done as long as the centroids move a considerable amount. After some iterations, the movements will fall below a threshold and we consider clustering to be converged.

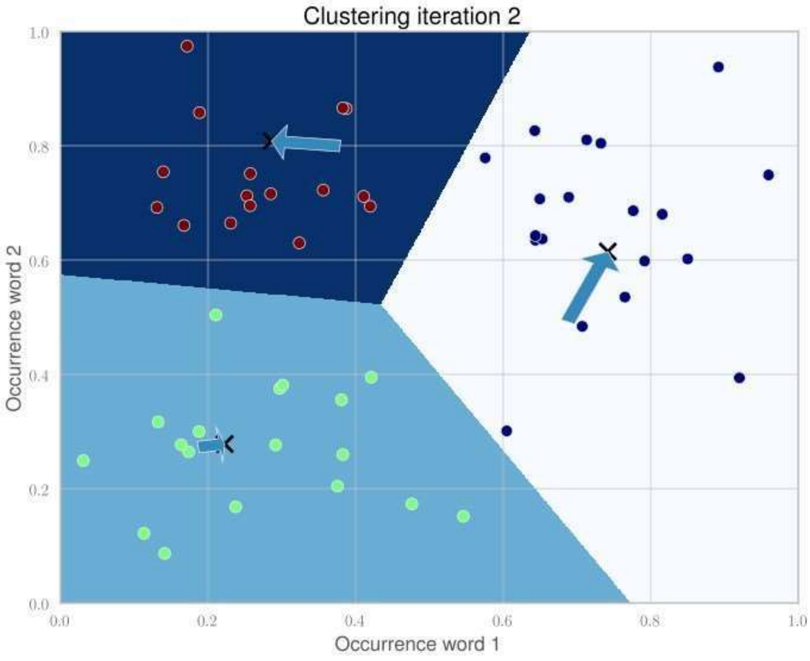
Let us play this through with a toy example of posts containing only two words. Each point in the following chart represents one post:



After running one iteration of KMeans, that is, taking any two vectors as starting points, assigning labels to the rest, and updating the cluster centers to be the new center point of all points in that cluster, we get the following clustering:



Because the cluster centers are moved, we have to reassign the cluster labels and recalculate the cluster centers. After iteration 2, we get the following clustering:



The arrows show the movements of the cluster centres. After five iterations in this example, the cluster centres don't move noticeably any more (Scikit's tolerance threshold is 0.0001 by default). Now we will generate the recommendation/related post

After the clustering has settled, we just need to note down the cluster centres and their identity. When each new post comes in, we have to vectorise and compare it with all the cluster centres. The cluster center with the smallest distance to our new post vector belongs to the cluster we will assign to the new post.

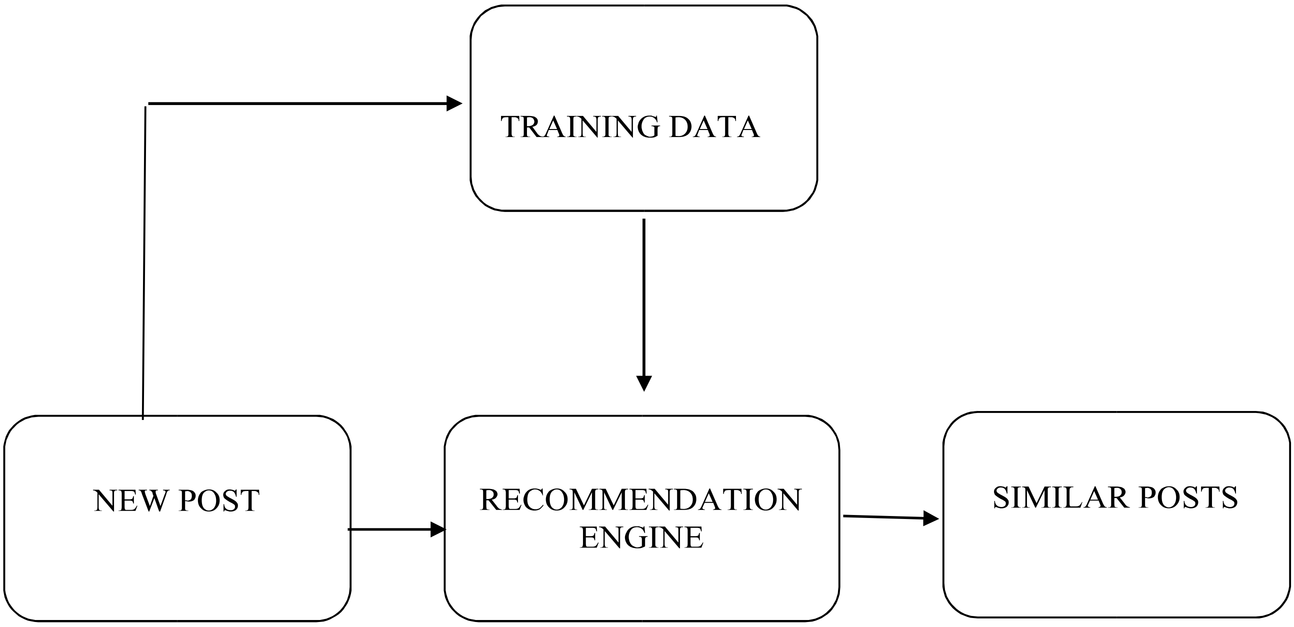


Fig. Flowchart for generating similar post.

```
>>>dist = sp.linalg.norm((new_post_vec - vectorized[i]).toarray())
```

The vectorizer which will be used will convert our posts to ti-idf weight matrix, since there will be increasing number of posts, the above distance calculation method will take more time. Therefore, to reduce the time and to reduce the load at the server we must convert these matrices to one dimensional array before calculating distance with the help of same toarray() function i.e.

```
>>>new_post_Vec.toarray()
>>>vectorized[i].toarray()
>>> dist = sp.linalg.norm(new_post_vec - vectorized[i])
```

What we have accomplished after vectorising the post?

1. We have fitted each post in their respective cluster using K-means clustering, where we have initialized the num_cluster i.e. number of clusters.

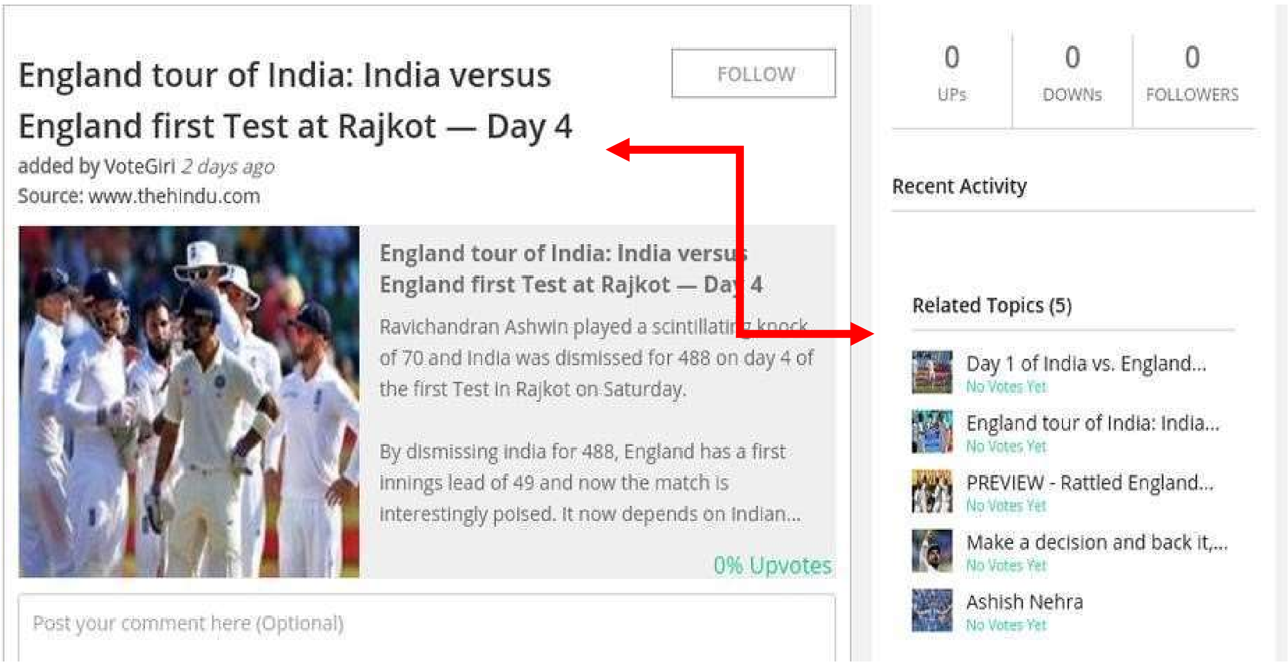
2. We have also initialized the no. Of iterations for a particular cluster.

3. Then, on the arrival of new post we have vectorised that post so that we can fit it into respective cluster after predicting with km.predict

4. Then for the new post, we compared distances from each post so that the posts with smallest distance can be recommended.

CHAPTER 6
RESULT

Our Recommendation system generates similar posts with respect to current post, which can be seen under Related Topics banner as shown below:



England tour of India: India versus England first Test at Rajkot — Day 4

FOLLOW

added by VoteGiri *2 days ago*

Source: www.thehindu.com

England tour of India: India versus England first Test at Rajkot — Day 4

Ravichandran Ashwin played a scintillating knock of 70 and India was dismissed for 488 on day 4 of the first Test in Rajkot on Saturday.

By dismissing India for 488, England has a first innings lead of 49 and now the match is interestingly poised. It now depends on Indian...

0% Upvotes

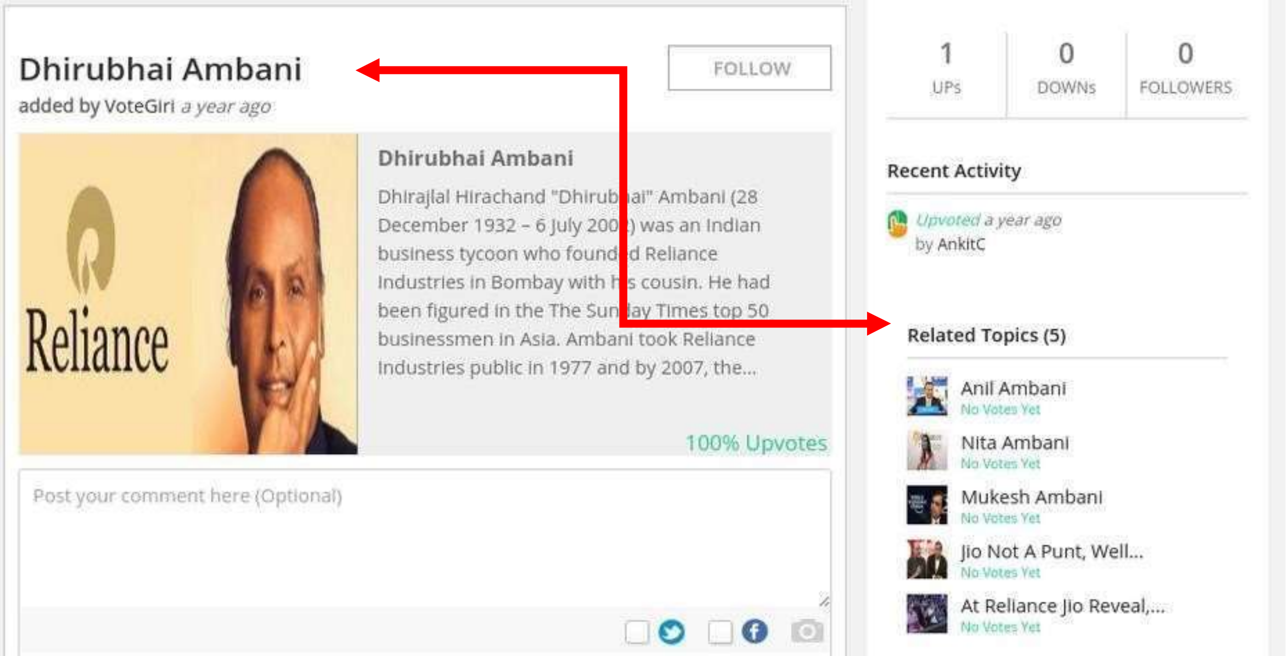
Post your comment here (Optional)

0	0	0
UPs	DOWNs	FOLLOWERS

Recent Activity

Related Topics (5)

- Day 1 of India vs. England... No Votes Yet
- England tour of India: India... No Votes Yet
- PREVIEW - Rattled England... No Votes Yet
- Make a decision and back it,... No Votes Yet
- Ashish Nehra No Votes Yet



Dhirubhai Ambani

FOLLOW

added by VoteGiri *a year ago*

Dhirubhai Ambani

Dhirajlal Hirachand "Dhirubhai" Ambani (28 December 1932 – 6 July 2002) was an Indian business tycoon who founded Reliance Industries in Bombay with his cousin. He had been figured in the The Sunday Times top 50 businessmen in Asia. Ambani took Reliance Industries public in 1977 and by 2007, the...

100% Upvotes

Post your comment here (Optional)

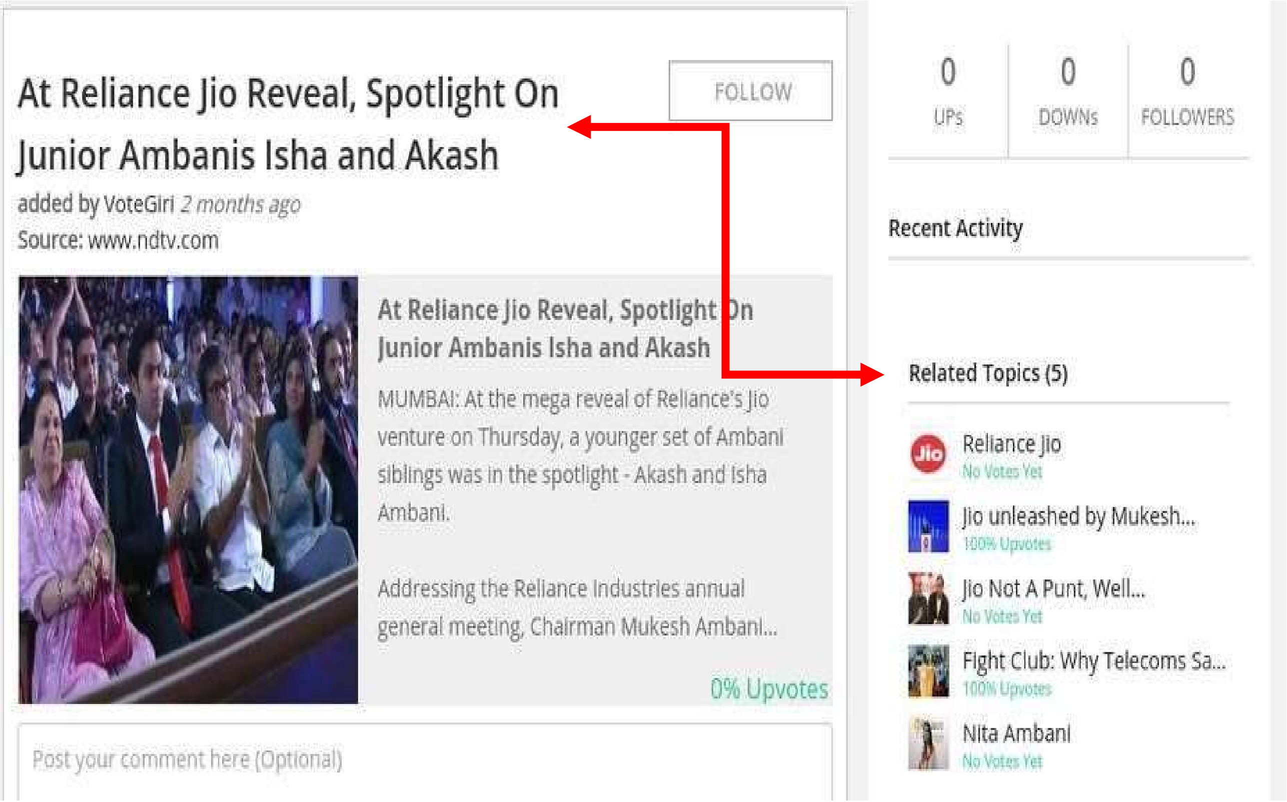
1	0	0
UPs	DOWNs	FOLLOWERS

Recent Activity

Upvoted a year ago by AnkitC

Related Topics (5)

- Anil Ambani No Votes Yet
- Nita Ambani No Votes Yet
- Mukesh Ambani No Votes Yet
- Jio Not A Punt, Well... No Votes Yet
- At Reliance Jio Reveal,... No Votes Yet



At Reliance Jio Reveal, Spotlight On Junior Ambanis Isha and Akash

FOLLOW

added by VoteGiri *2 months ago*

Source: www.ndtv.com

At Reliance Jio Reveal, Spotlight On Junior Ambanis Isha and Akash

MUMBAI: At the mega reveal of Reliance's Jio venture on Thursday, a younger set of Ambani siblings was in the spotlight - Akash and Isha Ambani.

Addressing the Reliance Industries annual general meeting, Chairman Mukesh Ambani...

0% Upvotes

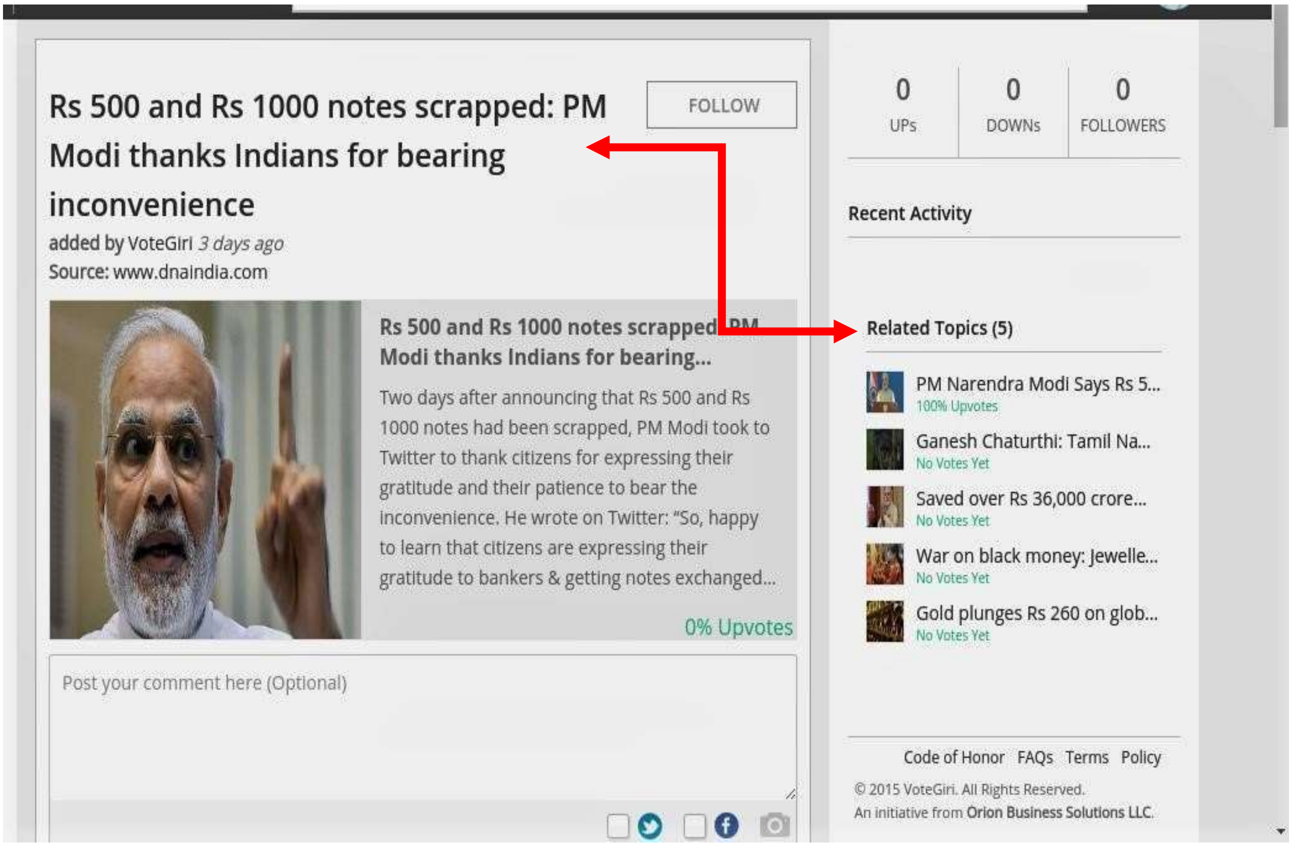
Post your comment here (Optional)

0	0	0
UPs	DOWNs	FOLLOWERS

Recent Activity

Related Topics (5)

- Reliance Jio — No Votes Yet
- Jio unleashed by Mukesh... — 100% Upvotes
- Jio Not A Punt, Well... — No Votes Yet
- Fight Club: Why Telecoms Sa... — 100% Upvotes
- Nita Ambani — No Votes Yet



Rs 500 and Rs 1000 notes scrapped: PM Modi thanks Indians for bearing inconvenience

FOLLOW

added by VoteGiri *3 days ago*

Source: www.dnaindia.com

Rs 500 and Rs 1000 notes scrapped: PM Modi thanks Indians for bearing...

Two days after announcing that Rs 500 and Rs 1000 notes had been scrapped, PM Modi took to Twitter to thank citizens for expressing their gratitude and their patience to bear the inconvenience. He wrote on Twitter: "So, happy to learn that citizens are expressing their gratitude to bankers & getting notes exchanged...

0% Upvotes

Post your comment here (Optional)

0	0	0
UPs	DOWNs	FOLLOWERS

Recent Activity

Related Topics (5)

- PM Narendra Modi Says Rs 5... — 100% Upvotes
- Ganesh Chaturthi: Tamil Na... — No Votes Yet
- Saved over Rs 36,000 crore... — No Votes Yet
- War on black money: Jewelle... — No Votes Yet
- Gold plunges Rs 260 on glob... — No Votes Yet

Code of Honor FAQs Terms Policy

© 2015 VoteGiri. All Rights Reserved.

An initiative from **Orion Business Solutions LLC.**

Conclusion

The internship is a bridge between the theoretical knowledge and the practical or the reality work. It is to explore a career interest, develop skills, and gain experience. It is up to us to make the most of the internship by learning as much as we can.

Transitioning from college to career can be intimidating. Many graduates struggle to identify the right career path or find themselves getting lost in the competitive job market, this is where internships helps students also employers increasingly hire students with internship experience over students who have none.

At Orion Business Solutions, Indore, you will be provide with an opportunity to work in the different part of the project work which helped me gain more knowledge by seeing what they work in their own office and what is their main responsibilities to the client and also to each other?

However, this internship program was not free from challenges. We faced many challenges since we were working on recommendations engine, because there is no particular path to develop a recommendation engine. Our primary challenge is the data. Since, there is no Analytic Sandbox is present for the company we needed to get permissions for the sample or the authentic data from seniors. For recommendations systems we must strive for best recommended solutions and thus this was one of the most difficult challenge for us to overcome, another most difficult challenge for us is to optimize the program so that it would take less time which will in turn reduce the load at the server. The challenges we faced are interesting to go through, as they bring the best in us.

Overall this internship program laid sound foundation for us to start our career.

References

1. Data Science & Big Data Analytics Book , EMC(pp. 26 – 62 , 256 – 291).
2. Willi Richert &Luis Pedro Coelho, Building Machine Learning Systems with Python [Book]
3. Pandas Documentation [http://pandas.pydata.org/pandas-docs/stable/].
4. Scipy and Numpy Documentation [https://docs.scipy.org/doc/numpy/reference/generated/numpy.linalg.norm.html]
5. Mysql Connector [https://dev.mysql.com/doc/connector-python/en/connector-pythonexample-cursor-select.html]
6. Natural Language Toolkit — NLTK 3.0 documentation [www.nltk.org/]