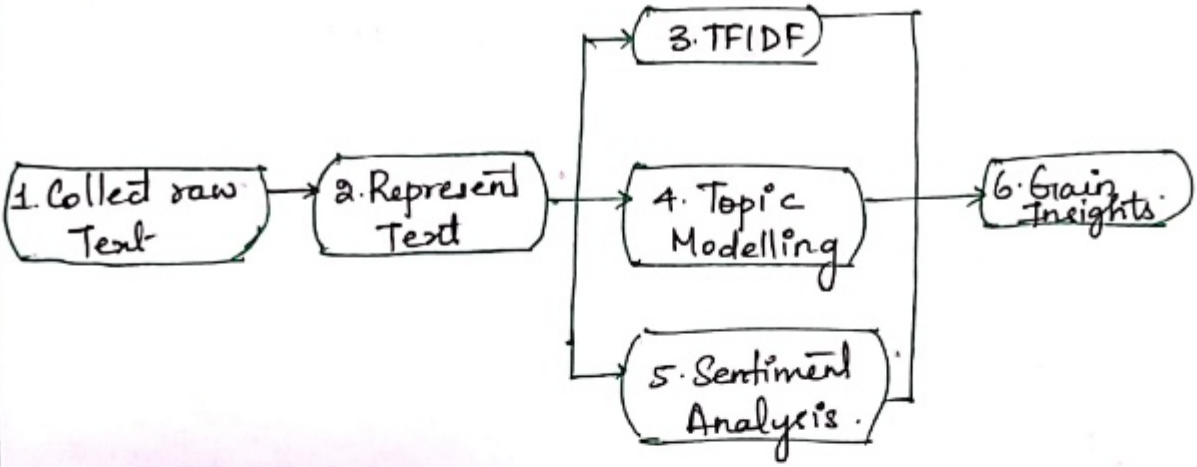
Semester : VI Subject : DAV Academic Year: 2023- 2024

Text Analysis Example :

Consider any example of a product it has to undergoe the following steps. Consider the company ACME, maker of 100 products : bPhone and bEbook.

It needs to answer questions such as these.

Are people mentioning its products?
What is being said? Are the products seen as good or bad? If people think an ACME product is bad, why? For example, are they complaining about the battery life of the bPhone, or the response time in their Ebook?



1 Collect Raw Data :-

* It includes the phase 1 and phase 2 of Data Analytics Life cycle.
* The Data Science team investigates the problem, understands the necessary data sources, and formulates initial hypothesis.

Semester: VI Subject: DAV Academic Year: 2023-2024

* The Data Science teams starts by actively monitoring various websites for user-generated contents.
* The user-generated contents being collected could be related articles from news portals and blogs, comments on ACME's products from online shops or review sites, or social media posts that contain keywords bPhone or Ebook.
* They gather data through semi-structured data - XML, HTML, RSS feeds, JSON etc.

Example:
A sample tweet that contains the keyword bPhone:

```
01 {
02 {"created_at": "Thu Aug 15 20:06:48 +0000 2013",
03 "coordinates": {
04   "type": "Point",
05   "coordinates": [.
 :        :
22 "text": "I once had a gf back in the day. Then the bPhone
            came out lol."
```

In this example Fields created_at at line 2 and text at line 22 in the previous tweet provide the information that interests ACME. The created_at entry stores the timestamp that the event was published, and the text field stores the main content of the Twitter post.

Semester : VI Subject : DAV Academic Year: 2023-2024

Many news portal and blogs provide data feeds that are in a open standard format, such as RSS and XML. As an example, an RSS feed for a phone review blog is shown next.

```
01 <channel>
02   <title> All about bPhone </title>
03   <description> My phone Review site </description>
04   <link> http://www.phones.com/link.html </link>.
05
06 <item>
07    <title> bPhone : The best </title>.
08    <description> I love my bPhone! </description>
09    <link> http://www.phones.com/link.html </link>.
10.   <guid isPermaLink = "false"> 1102345 </guid>
11.   <pubDate> Tue 29 Aug 2011 09:00:00 -400 </pubDate>
12. <item>
13. </channel>.
```

The content from the title (line 7), the description (line 8), and the published date (pubDate, line 11) is what ACME is interested in.

If the plan is to collect user comments on ACME's products from online shops and review sites where API's or data feeds are not provided, the team may have to write web scrapers to parse webpages and automatically extract the interesting data from those HTML files.