**Summary Report**

**Patent Data Extraction**

**- Phase 1**

**Team Kick SAS**

**Members:**

**Atluri Laxmi Narayana 800907065**

**Mohan Kumar Nallapareddy 800886720**

**Punniya Dharshan Ganesan 800894568**

**Reethu Gundavarapu 800903616**

**Sharan Kumar Reddy Kadire 800888186**

**Sri Harsha Degala 800902540**

**Executive Summary- Patent Analysis for Microsoft**

**Microsoft Corporation** is an American multinational technology company headquartered in Redmond, Washington, that develops, manufactures, licenses, supports and sells computer software, consumer electronics and personal computers and services. Its best known software products are the Microsoft Windows line of operating systems, Microsoft Office suite, and [Internet Explorer](https://en.wikipedia.org/wiki/Internet_Explorer) and [Edge](https://en.wikipedia.org/wiki/Microsoft_Edge) [web browsers](https://en.wikipedia.org/wiki/Web_browser). Its flagship hardware products are the [Xbox](https://en.wikipedia.org/wiki/Xbox) game consoles and the [Microsoft Surface](https://en.wikipedia.org/wiki/Microsoft_Surface) tablet lineup. It is the [world's largest software maker](https://en.wikipedia.org/wiki/List_of_the_largest_software_companies) measured by revenues. It is also one of the [world's most valuable companies](https://en.wikipedia.org/wiki/List_of_corporations_by_market_capitalization).

Microsoft was founded by Paul Allen and Bill Gates on April 4, 1975, to develop and sell [BASIC](https://en.wikipedia.org/wiki/BASIC) [interpreters](https://en.wikipedia.org/wiki/Interpreter_(computer_software)) for [Altair 8800](https://en.wikipedia.org/wiki/Altair_8800). It rose to dominate the [personal computer](https://en.wikipedia.org/wiki/Personal_computer) [operating system](https://en.wikipedia.org/wiki/Operating_system) market with [MS-DOS](https://en.wikipedia.org/wiki/MS-DOS) in the mid-1980s, followed by [Microsoft Windows](https://en.wikipedia.org/wiki/Microsoft_Windows) The company's 1986 [initial public offering](https://en.wikipedia.org/wiki/Initial_public_offering), and subsequent rise in its share price, created three billionaires and an estimated 12,000 millionaires from Microsoft employees. Since the 1990s, it has increasingly diversified from the operating system market and has [made a number of corporate acquisitions](https://en.wikipedia.org/wiki/List_of_mergers_and_acquisitions_by_Microsoft). In May 2011, Microsoft acquired [Skype Technologies](https://en.wikipedia.org/wiki/Skype_Technologies) for $8.5 billion in its largest acquisition to date. (Source: Wikipedia)

We analyzed past 10 years data on Microsoft patents filed and was able to achieve some patterns in their patent filing. We also performed text analytics on the data extracted from these files. We used Hive, Hadoop, R and Tableau for data extraction, analysis and visualization.

Top three patents:

US8094129- Touch sensing using shadow and reflective modes.

US8082517- Application sharing viewer presentation.

US853340- Accelerating parallel transactions using cache resident transactions.

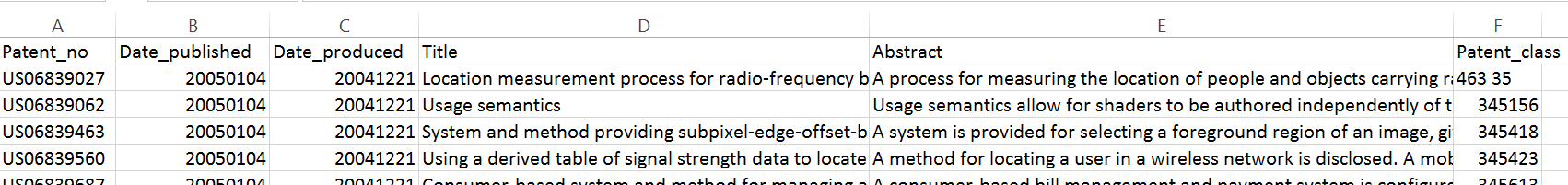
**Data Extraction:**

* We took the data from the patentbibilio folder and unzipped the files.
* We used Shell script to unzip each and every file and copy the **Microsoft** patents to an xml file.



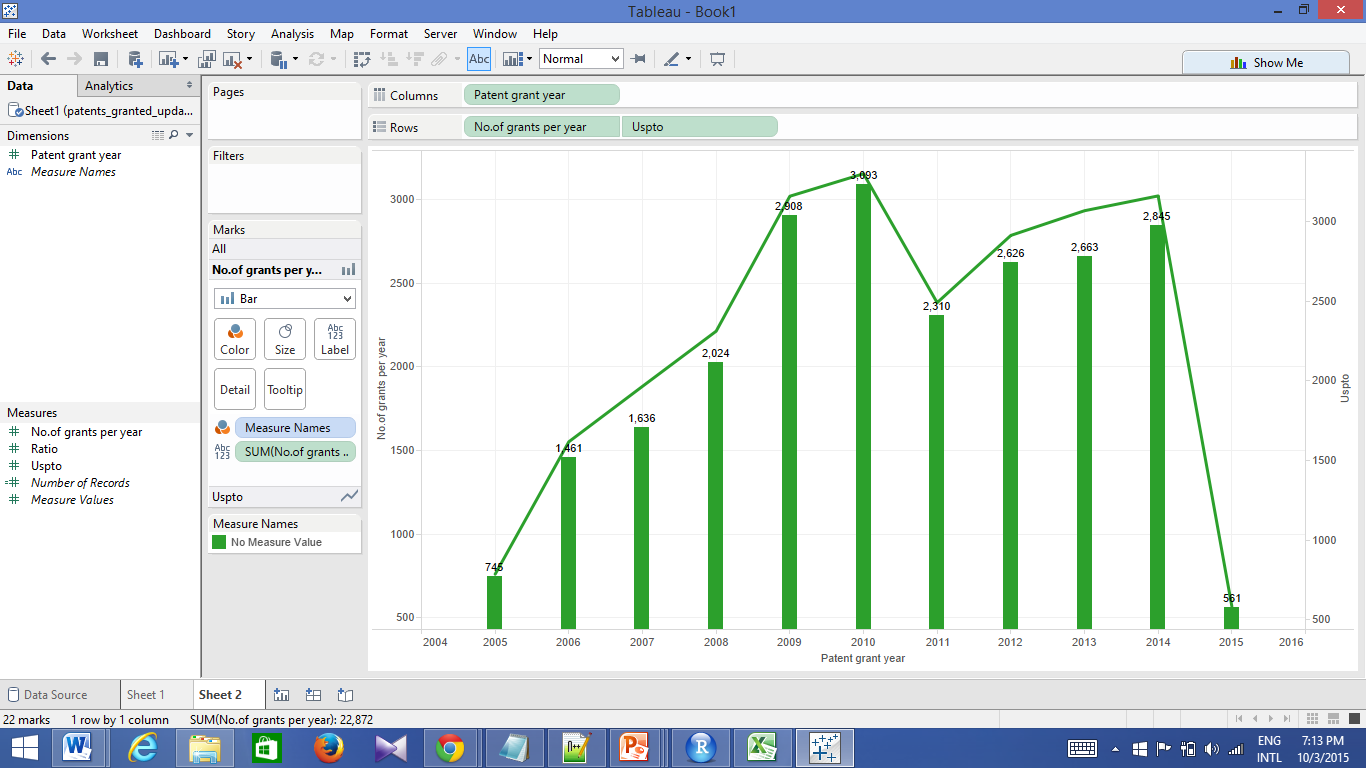
**Data Wrangling:**

* We initially tried XML Parsing and Map Reduce code for data wrangling but we faced few challenges in our execution-
  + Xml parsing in java could not fetch required results as our xml files were not in pure xml format.
  + The mapper code was not able to execute huge data, it goes to hung state.
  + It returned junk values in the results.
* Due to the above challenges , we decided to go for Hive.

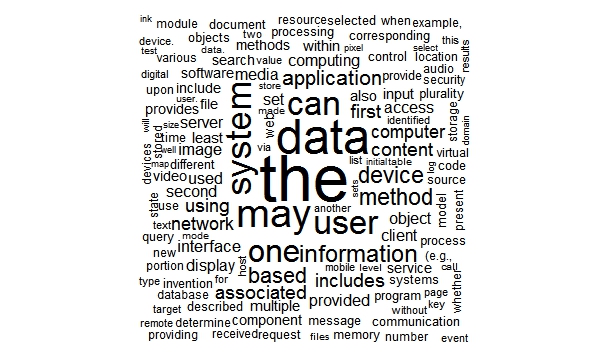


|  |  |  |  |
| --- | --- | --- | --- |
| Patent Grant year | No of Grants identified | USPTO | Ratio |
| 2005 | 745 | 787 | 0.946633 |
| 2006 | 1,461 | 1,618 | 0.902967 |
| 2007 | 1,636 | 1,966 | 0.832146 |
| 2008 | 2,024 | 2,312 | 0.875433 |
| 2009 | 2,908 | 3,159 | 0.920544 |
| 2010 | 3,093 | 3,300 | 0.937273 |
| 2011 | 2,310 | 2,491 | 0.927338 |
| 2012 | 2,626 | 2,913 | 0.901476 |
| 2013 | 2,663 | 3,068 | 0.867992 |
| 2014 | 2,845 | 3,161 | 0.900032 |
| 2015 | 561 | 586 | 0.957338 |

***Ratio of the no of patents identified to the total no of patents filed per year***

**

***The visualization of patents extracted versus the USPTO patents.***



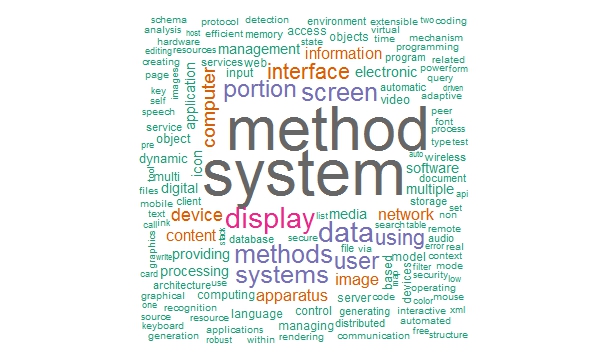
***Word cloud on initial data:***

The stop words like ‘the’ are occupying more space of the data and hence such type of words are more space in the above word cloud.



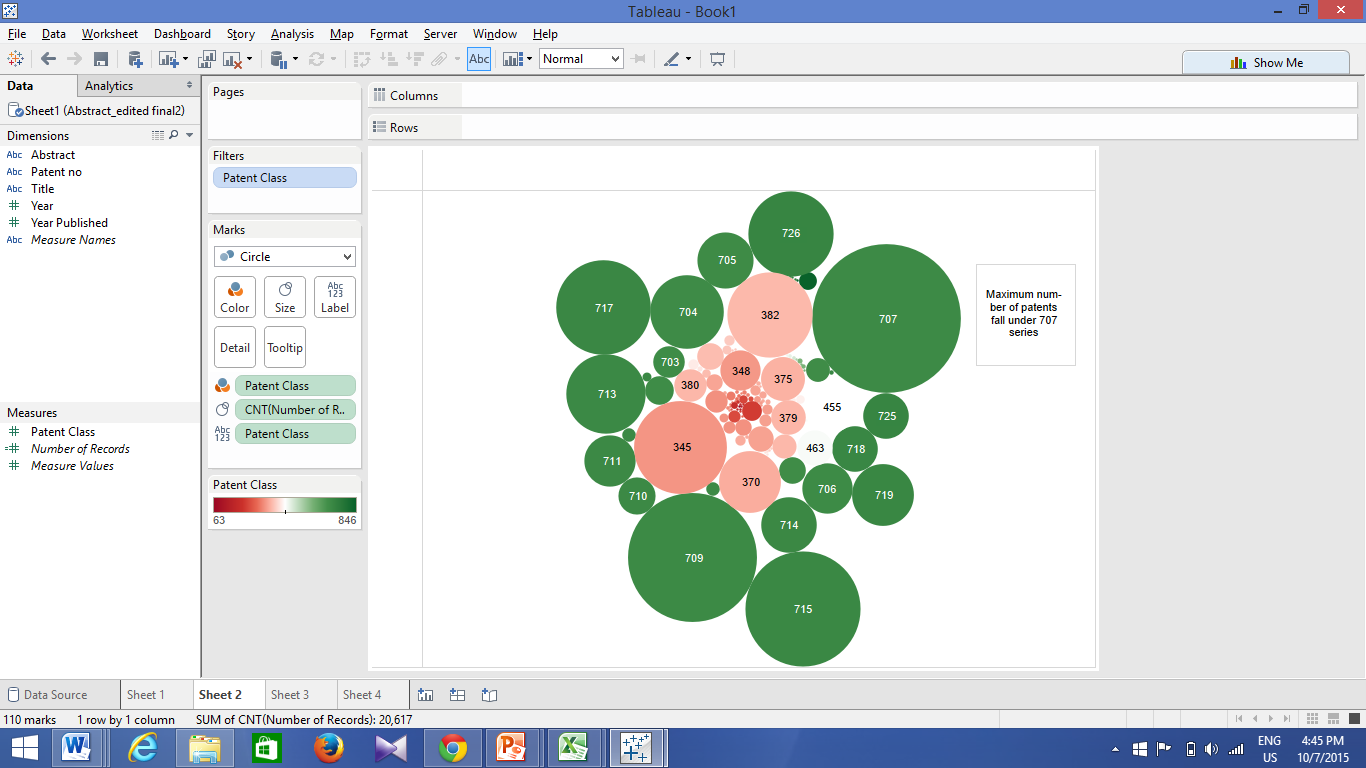
***Word cloud after running stop words and adding colour:***

After removing the stop words from the data, the above word cloud is appearing and data is the most used word according to the above word cloud.



***Word cloud for titles alone with colour***

The above word cloud depicts the usage of the words in titles of the patents.



***Patent class grouped by count***