**Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR Nagar, MSRIT post, Bangalore-54

A Dissertation Report on

‘Fuzzy String Matching of Financial Data using concepts of Fuzzy Logic’

Submitted by

**Rishabh Malhotra** 1MS15CS103

**Saumit C** 1MS15CS109

**B.Sandeep Reddy** 1MS15CS157

*Bachelor of Engineering in Computer Science & Engineering*

Under the guidance of

**Name of the guide** : Ms. Sowmya B J

**Designation** : Assistant Professor

**Dept name** : Computer Science & Engg

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

www.msrit.edu, **20177**

1. **Introduction:**

**Fuzzy logic** is a form of many-valued logic in which the truth values of variables may be any real number between 0 and 1. It is employed to handle the concept of partial truth, where the truth value may range between completely true and completely false.By contrast, in Boolean logic, the truth values of variables may only be the integer values 0 or 1. Furthermore, when linguistic variables are used, these degrees may be managed by specific (membership) functions.

The term **fuzzy logic**was introduced with the 1965 proposal of fuzzy set theory by Lotfi Zadeh.

Fuzzy logic had however been studied since the 1920s, as infinite-valued logic—notably by Łukasiewicz and Tarski.

Classical logic only permits conclusions which are either true or false. However, there are also propositions with variable answers, such as one might find when asking a group of people to identify a color. In such instances, the truth appears as the result of reasoning from inexact or partial knowledge in which the sampled answers are mapped on a spectrum.

Humans and animals often operate using fuzzy evaluations in many everyday situations. In the case where someone is tossing an object into a container from a distance, the person does not compute exact values for the object weight, density, distance, direction, container height and width, and air resistance to determine the force and angle to toss the object. Instead the person instinctively applies quick "fuzzy" estimates, based upon previous experience, to determine what output values of force, direction and vertical angle to use to make the toss.

Both degrees of truth and probabilities range between 0 and 1 and hence may seem similar at first, but fuzzy logic uses degrees of truth as a mathematical model of vagueness, while probability is a mathematical model of ignore.

Fuzzy logic seems closer to the way our brains work. We aggregate data and form a number of partial truths which we aggregate further into higher truths which in turn, when certain thresholds are exceeded, cause certain further results such as motor reaction. A similar kind of process is used in neural networks, expert systems and other [artificial intelligence](http://searchcio.techtarget.com/definition/AI) applications. Fuzzy logic is essential to the development of human-like capabilities for AI, sometimes referred to as artificial general intelligence: the representation of generalized human cognitive abilities in software so that, faced with an unfamiliar task, the AI system could find a solution.

For example, - A restaurant’s rating system could make use great of fuzzy logic while calculating tips. The latter’s ratings would not be completely 0 (worst) or 10 (excellent), but would lie in between the two.

1. **Data Set Description:**

There are two datasets being used for the purpose of ‘Fuzzy String Matching’, the first one consists of all the S&P 500 companies listed on the New York Stock Exchange.

However, the second dataset contains all the publically listed companies on the NYSE.

The first dataset contains two attributes namely, the names of all companies and their average stock volumes (S&P Volumes) . **It contains about 505 tuples**.

The second dataset contains three attributes namely, names of all companies, the sector they represent and the industry they are classified into. **It contains around 3500 tuples**.

(i) **Source of Dataset:**

First Dataset: <http://www.princeton.edu/~otorres/sandp500.csv>

(The website of Princeton University)

Second Dataset: <http://www.nasdaq.com/screening/companies-bindustry.aspx?exchange=NYSE>

(The Official Website of Nasdaq, NYSE)

(ii) **Attributes Description:**

First Dataset:

1. S&P Listed Company Names:

These are the top 500 companies on the NYSE, these companies compute the Nasdaq and are an indicator of the economy. They range from various sectors such as Manufacturing, F&B, Oil, Durables & Non-Durables etc.

1. Volumes of the stock:

The values mentioned in the volumes attribute define the stock volumes (i.e, the number of stocks held by public shareholders).

Second Dataset:

1. Names of all companies listed on NYSE:

The list contains all the publically listed companies on the NYSE, including all S&P listed companies.

1. Sector:

The sector attribute represents the sector the stocks represent.

1. Industry:

This attributes is the sub-attribute of sector, it represents which industry in which sector the stock represents.

(iii) **Dataset Size:**

Dataset 1 : 16KB

Dataset 2 : 181KB

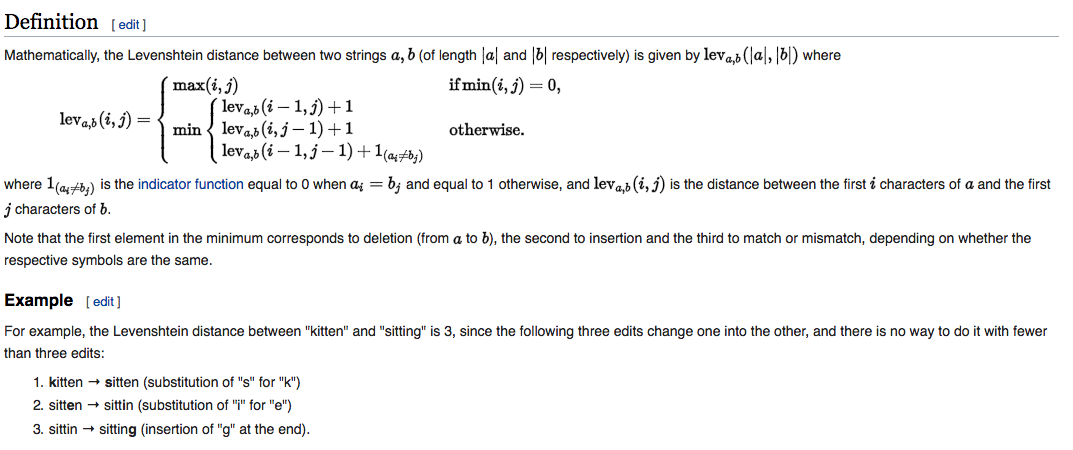
(iv) **Inferences Drawn:**

Dataset 1 : The attributes show us the public shareholding of S&P stocks on NYSE.

Dataset 2 : The attributes tell us about all the stocks listed on NYSE and their sectors &

industry.

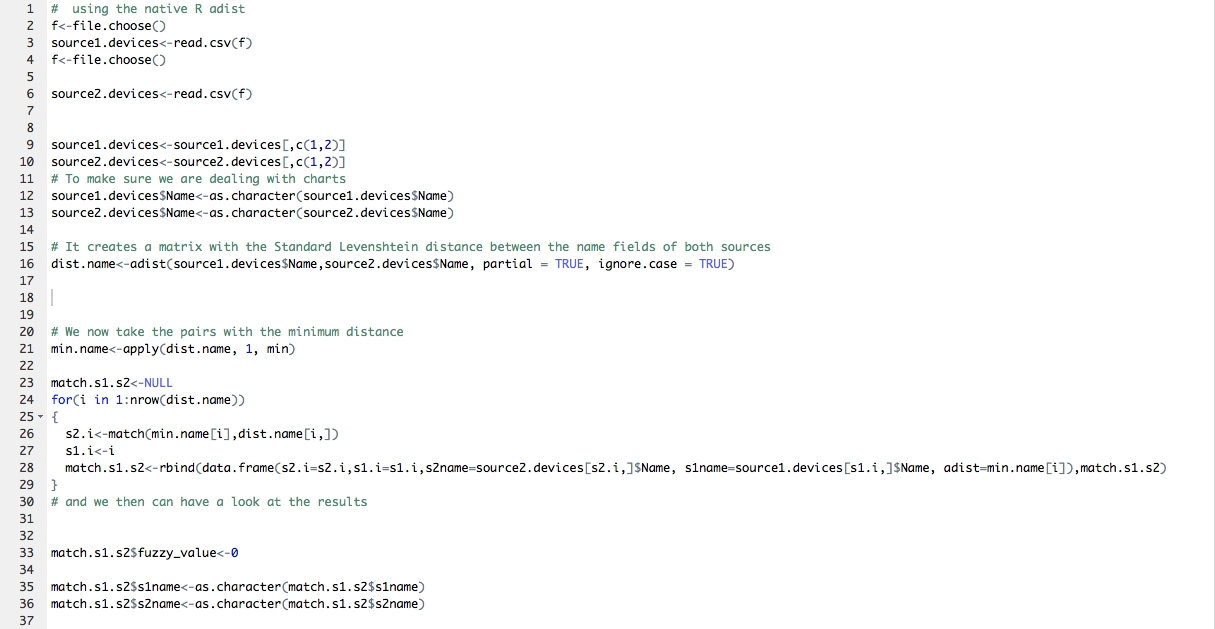
1. **Algorithm Description:**

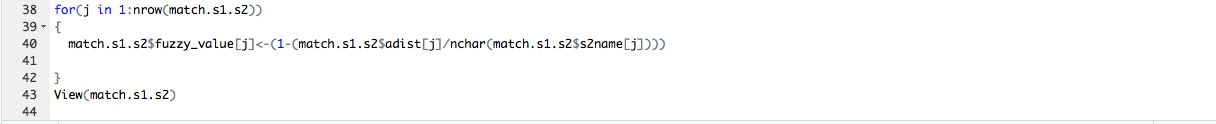


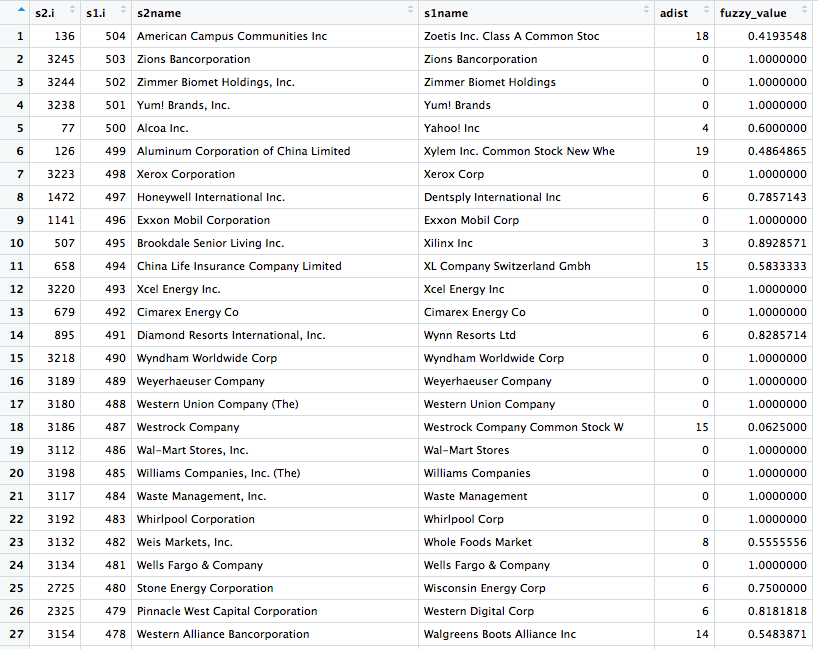
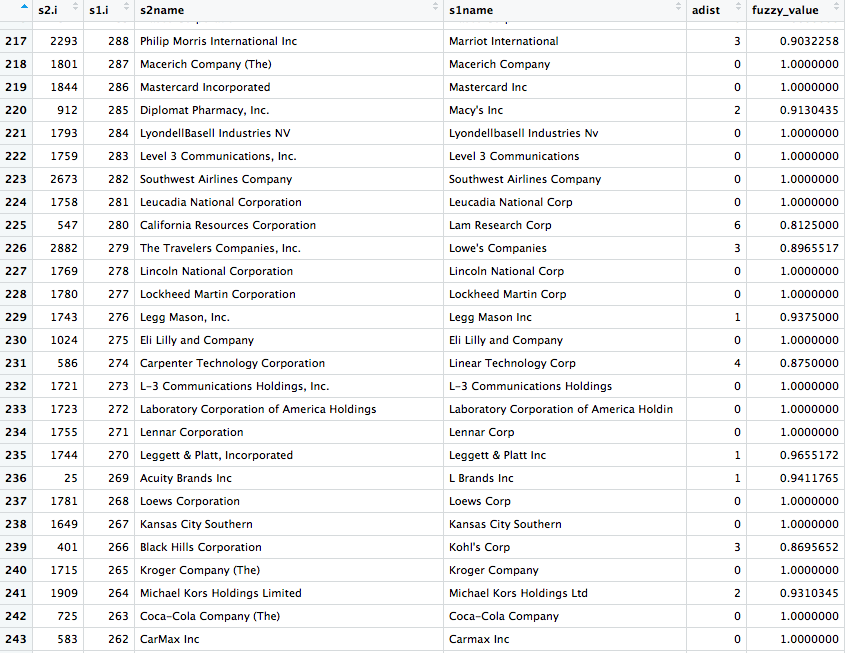
Approximate String Matching or Fuzzy String Matching:

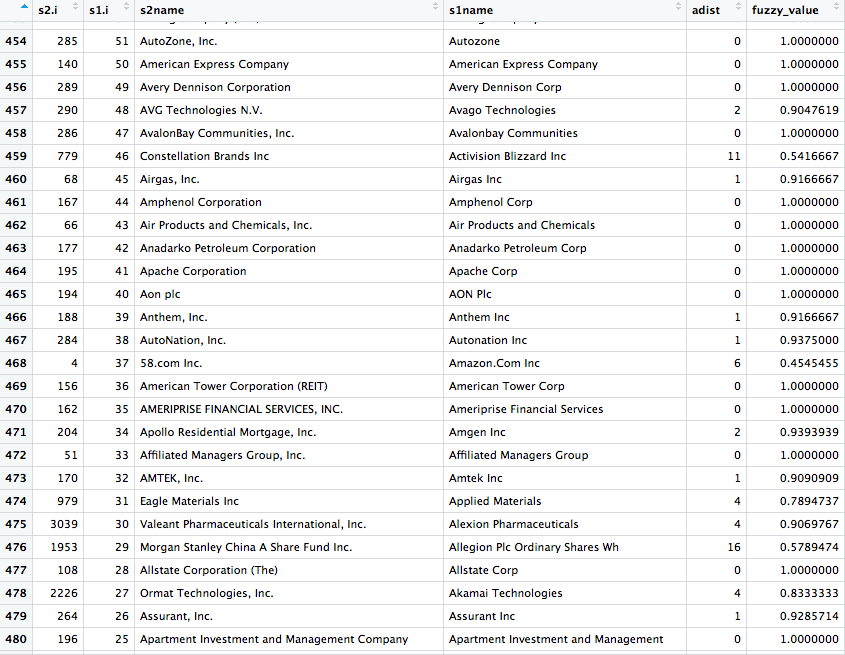
**Approximate string matching** (often colloquially referred to as **fuzzy string searching**) is the technique of finding strings that match a pattern approximately (rather than exactly). The problem of approximate string matching is typically divided into two sub-problems: finding approximate substring matches inside a given string and finding dictionary strings that match the pattern approximately.

1. **Snapshot of code:**





1. **Result Snapshots:** ****



**Resultset Description:**

From the above results, it is clearly seen that companies listed on the S&P 500 were also present on the NYSE,

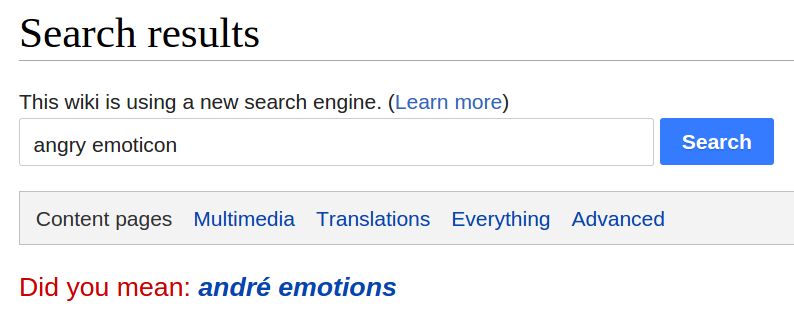
We use the adist function which computes the string distance, as seen in the algorithm section of this report.

To compute the fuzzy value, the adist value is divided by the string length and the this entity is subtracted by 1.

The results above were established using concepts of Fuzzy String Matching where if the string is present in both, we get a non-zero value, however for an exact match the value would be exactly 1.

If no match is obtained, then the value is 0.

1. **Social Impact:**
2. This technique guides equity investors on investing in the correct stock mentioned in the S&P 500 list, which also in turn acts like a good indicator of the U.S Economy, hence strengthing the US Economy.



This can also be used in search engines to predict what information the user wishes to get and if the user makes a silly mistake we can get to know using the fuzzy string matching algorithm.

(iii) With the availability of large amounts of DNA data, matching of nucleotide sequences has become an important application. Approximate matching is also used in spam filtering. String matching cannot be used for most binary data, such as images and music. They require different algorithms, such as acoustic fingerprinting.