# Title: Fuzzy de-duplication of leads and contacts - May 2019

### Feature

Build scalable framework which can identify duplicates in spite of spelling mistakes, abbreviations, fields mismatch, phonetic similarity, etc.

### **Team**

| Team member name | Role |
| --- | --- |
| Bharathi | Freddy - Engineering |
| Srivatsa | Freddy - Data Science |
| Suvrat | Freddy - Lead |
| Ajeet | Core engineer |
| **Logesh** | **QA** |
| Sivalingam | Core lead |
| Madhav, Sivakumar | Devops |
| Swaminathan | Freddy |
| Sudharshan | Architect |
| Aditi, Srivatsan | Product |

### **Timeline**

|  |  |
| --- | --- |
| 1st to 23rd April 2019 | Deployment on docker/staging and testing |
| 23rd to 30 April 2019 | Production setup, offline scripts |
| 2nd - 9th May 2019 | Testing on Prod accounts. Some issues were  fixed by core team. |
| 11th May 2019 - Ongoing | Live on Freshdesk. Metric calculation/Monitoring stats. |

### **Success Metric to be measured**

Online and offline metric can be found [here](https://docs.google.com/spreadsheets/d/1dxdYQk1EiMQy1gnYLQ5FbnZ6T7EqY_ksIvsmyzPzIko/edit?usp=sharing).

Specific QA test cases can be found [here](https://docs.google.com/spreadsheets/d/1ebsaVc2Xf-AL5cYVbx4Q891ChlOmkXJmaGjxNAkM9HI/edit?usp=sharing).

| Capability | ES based | Freddy based |
| --- | --- | --- |
| Phonetic match | No | Yes |
| Field mismatch | No | Yes |
| Abbreviation | No | Yes |
| Spelling mistake | No | Yes |
| Domain match | No | Yes |
| Smart ranking | No | Yes |
| Pure string match on name | Yes | No |

### Patent

We are working with the legal team for patent. Ticket was raised on 25th April for same.

### **Coming up**

Foreign languages support ( Non unicode characters )

Account/Company name deduplication

### **Apendix**

### **Architecture**

### 

### **Data science Model**

| Feature Name | Description | Why this feature is included ? | Priority Score |
| --- | --- | --- | --- |
| 1. Concat Name - Max Jaro Winkler | maximum of jaro winkler distances of (p\_fn + p\_ln, s\_fn+s\_ln) , (p\_ln+p\_fn, s\_fn+s\_ln) , (p\_fn + p\_ln, s\_ln+s\_fn) , (p\_ln + p\_fn, s\_ln+s\_fn) | Incorporates the similarity between the names of a pair of records, robust towards instances where the first name and last name are swapped |  |
| 2. Phone Match | exact match between (Last 6 digits of p\_mobile\_number , Last 6 digits of p\_work\_number) and (Last 6 digits of s\_mobile\_number, Last 6 digits of s\_work\_number) | robust towards variations in the way the phone numbers are entered - special characters, area and country codes at the beginning of the phone number | 13.82% |
| 3. Phone Given | if (p\_work\_number & p\_mobile\_number) or (s\_work\_number & s\_mobile\_number), return 1  else return 0 |  | 1.41% |
| 4. Company Name - Lname - Max Jaro Winkler | maximum of jaro winkler distances of (p\_company\_name, s\_company\_name) , (p\_company\_name , s\_last\_name) , (s\_company\_name, p\_last\_name) | captures similarity between company names or between company name and last name   * if two records are similar, then they might belong to the same company * also in some cases, the company name is entered in the last name field | 9.9% |
| 5. Number of exact token matches | intersection of no. of tokens / union of no. of tokens | captures similarity between two records based on the number of exact token matches - if this number is high, then there is a probable chance the two records are similar |  |
| 6. Number of phonetic token matches for first name and last name | intersection of no. of tokens / union of no. of tokens | captures the phonetic similarity between two names - they might differ in their spelling but might sound the same. Ex - John Smith & Jon Smyth |  |
| 7. Final Concat Name | Concat name + phonetic token match score | reinforces the concat score if the two names sound similar (Feature 6 and Feature 1 are not used explicitly - instead they are combined to form Feature 7) | 38.68% |
| 8. Final token match | Token match score with abbreviations of company names also included as tokens | reinforces the token match score if the abbreviations of the company names are similar. Ex - IBM and International Business Machines - these two have the same abbreviations and hence help increase the token match score. (Feature 5 is used along with abbreviations to form Feature 8) | 32.78% |
| 9. p\_empty\_fn | if the first name of the primary record is empty or not |  | 0.004% |
| 10. s\_empty\_fn | if the first name of the secondary record is empty or not |  | 0.66% |
| 11. p\_empty\_company | if the company name of the primary record is empty or not |  | 1.46% |
| 12. p\_empty\_company | if the company name of the secondary record is empty or not |  | 1.22% |
| 13. p\_email\_company | if an email address if present in the company name of the primary record or not |  | 0.002% |
| 14. s\_email\_company | if an email address if present in the company name of the secondary record or not |  | 0.006% |
| 15. p\_email\_ln | if an email address if present in the last name of the secondary record or not |  | 0.013% |
| 16. s\_email\_ln | if an email address if present in the last name of the secondary record or not |  | 0.012% |

