Phonetic Anagrams Report

Yu Sheng

2018/3/21

Soundex and Metaphone are two phonetic algorithms which are mainly used to find similar sounding words in English. Words will be coded into an index based on their pronunciation. In Soundex algorithm, index is composed by four characters started by an alphabet and then followed by three numbers. What’s more, it mainly encodes consonants. Soundex algorithm was first developed by Robert C. Russell and Margaret King[[8]](https://en.wikipedia.org/wiki/Soundex" \l "cite_note-8) Odell in around 1920.[1] In 1930s, a variation of Soundex called American Soundex is used in census. This algorithm is very useful when searching for similar sounding but different spelling names despite of its purpose of encode surnames for use in censuses. For example, when trying to find "Smyth", it is very common to misspell it as “Smith”, then this algorithm will be used to give people some suggestion like “Do you mean Smyth? ”[2] Under the needs for matching similar sounding words more accurately, Metaphone is invented by Lawrence Philips in 1990.[3] It uses information about variations and inconsistencies in English spelling and pronunciation to improve the accuracy thus are suitable for most words not only the names. Metaphone codes are composed by only letters and length is not restricted. It provides the base for many spell checker[4]. When person spells a wrong word and a Metaphone code will be generated and then we can check the corresponding class of this code or similar code to see if there is a alternative spelling. Lawrence Philips also invented Double Metaphone later as to increase the applicable range to other languages. Moreover, in 2009 same person produced the third and most powerful version called Metaphone 3 which has 99% accuracy for English words [3].

I found for Soundex algorithm, there are 424 classes. I found many names are incorrectly grouped especially in large classes like: “S500”, 'SONYA', 'SHANNON' “L200”, 'Louise', 'Lucy', 'Lizzie'; “L500”, 'Loma', 'Lynne', 'Lana'. Three examples of names I found to be misclassified are: “C623”, 'Christine’, ’Christa’, ‘Chrystal’; “K623”, ‘Kristine’, ‘Krista’, ‘Kristal’.

Top 10 largest classes are:

|  |  |
| --- | --- |
| J500 | 'JENNIE', 'JEAN', 'JANE', 'JANIE', 'JUANA', 'JOHANNA', 'JUNE', 'JEANNE', 'JOHNNIE', 'JOAN', 'JIMMIE', 'JANNIE', 'JOHN', 'JENNY', 'JOANNA', 'JOHNIE', 'JEANIE', 'JAYNE', 'JOANNE', 'JOANN', 'JEANNINE', 'JEANE', 'JAMIE', 'JAN', 'JEANINE', 'JEANNIE', 'JONNIE', 'JANA', 'JANINE', 'JONI', 'JANNA', 'JOANIE', 'JANN', 'JAMI', 'JOHNNA', 'JEANA', 'JEANNA', 'JANEEN', 'JONNA', 'JAIME', 'JENNA', 'JAYME', 'JAIMIE', 'JENA', 'JAMMIE', 'JANAE', 'JANAY', 'JOANA' |
| L200 | 'LOUISE', 'LUCY', 'LOIS', 'LIZZIE', 'LESSIE', 'LOUISA', 'LUCIA', 'LUISA', 'LUZ', 'LUCIE', 'LIZA', 'LEXIE', 'LOYCE', 'LOUIS', 'LISA', 'LEIGH', 'LESA', 'LIZ', 'LEESA', 'LISE', 'LEISA', 'LESIA', 'LISSA', 'LISHA', 'LAKISHA', 'LAKEISHA', 'LAKESHA', 'LACEY', 'LACY', 'LAKESHIA', 'LACIE', 'LACI', 'LEXUS', 'LEXI' |
| L500 | 'LENA', 'LEONA', 'LEONE', 'LINNIE', 'LINA', 'LONA', 'LENNIE', 'LONNIE', 'LUNA', 'LINNEA', 'LYNN', 'LENNA', 'LOMA', 'LONIE', 'LYNNE', 'LANA', 'LYN', 'LEANNA', 'LEANNE', 'LUANNE', 'LUANA', 'LONNA', 'LUANN', 'LEANN', 'LOUANN', 'LEEANN', 'LIANA' |
| D500 | 'DONNA', 'DIANA', 'DENA', 'DONA', 'DONNIE', 'DIANE', 'DINA', 'DAWN', 'DANA', 'DEAN', 'DEANNA', 'DIANNE', 'DIANNA', 'DIANN', 'DIAN', 'DINAH', 'DEANNE', 'DANNA', 'DEANA', 'DEENA', 'DEANN', 'DENEEN', 'DIONNE', 'DAWNA', 'DAYNA', 'DEMI’ |
| S500 | 'SONIA', 'SAMMIE', 'SONYA', 'SHANNON', 'SHAWN', 'SHAUNA', 'SHAWNA', 'SHANA', 'SHANNA', 'SIMONE', 'SHANNAN', 'SHANON', 'SHANI', 'SUNNY', 'SHONNA', 'SHEENA', 'SHAINA', 'SHAYNA', 'SHANIA', 'SIMONA', 'SEMONA', 'SHYANNE', 'SEMONE' |
| T500 | 'TINA', 'TOMMIE', 'TENNIE', 'TENA', 'TONI', 'TANYA', 'TANA', 'TAMMY', 'TAMI', 'TONYA', 'TAMMIE', 'TEENA', 'TONIA', 'TAMMI', 'TANIA', 'TAMIE', 'TAWANA', 'TAWNYA', 'TAWANNA', 'TIANA', 'TIANNA', 'TAWNY' |
| A420 | 'ALICE', 'ALYCE', 'ALICIA', 'ALEXIS', 'ALISA', 'ALESIA', 'ALISHA', 'ALECIA', 'ALYSSA', 'ALISSA', 'ALYSIA', 'ALEXA', 'ALESHA', 'ALYCIA', 'ALYSHA', 'ALEXIA', 'ALYSE', 'ALEXUS', 'ALEX', 'ALLYSSA', 'ALYSA’ |
| C623 | 'CHRISTINE', 'CHRISTINA', 'CRYSTAL', 'CRISTINA', 'CHRISTENE', 'CHRISTEEN', 'CHRISTIE', 'CHRISTY', 'CHRISTI', 'CHRISTA', 'CHRYSTAL', 'CRISTY', 'CHRISTAL', 'CHRISTOPHER', 'CHRISTIAN', 'CHRISTEN', 'CHRISTEL', 'CHRISTIN', 'CRISTAL', 'CRISTIN', 'CHRISTIANA' |
| R500 | 'RENA', 'RAMONA', 'ROWENA', 'ROMA', 'RENEE', 'ROMAINE', 'RONNIE', 'RONNA', 'RENE', 'RENAE', 'ROMONA', 'RONI', 'RENEA', 'RONA', 'RYAN', 'RAINA', 'RHIANNON', 'REYNA', 'REINA' |
| C640 | 'CAROL', 'CHARLIE', 'CORAL', 'CAROLE', 'CREOLA', 'CARYL', 'CARROLL', 'CARLA', 'CHERYL', 'CAROLEE', 'CARROL', 'CARL', 'CHERYLE', 'CHARLA', 'CARLY', 'CARLEY', 'CARLIE', 'CARLI', 'CARLEE' |

I find for Metaphone algorithm, there are 456 classes. Three examples I found for

names incorrectly grouped are: “TR”, Drew, Tyra; “TN”, 'Dianna', 'Tawny'; SNT 'Santa', 'Cindy'. It is very hard to find misclassified examples in Metaphone algorithm, names in one group always have at least some part of similar sound, I find one is “XL”, ‘Chloe’ should be placed to KLT ['Claudie', 'Colette',…].

Top 10 largest classes are:

|  |  |
| --- | --- |
| JN | 'Jennie', 'Jean', 'Jane', 'Janie', 'Juana', 'June', 'Jeanne', 'Johnnie', 'Joan', 'Jannie', 'John', 'Jenny', 'Joanna', 'Johnie', 'Gene', 'Jeanie', 'Jayne', 'Joanne', 'Joann', 'Jeane', 'Jan', 'Jeannie', 'Jonnie', 'Jana', 'Gina', 'Joni', 'Janna', 'Ginny', 'Gena', 'Joanie', 'Jann', 'Johnna', 'Jeana', 'Jeanna', 'Jonna', 'Jenna', 'Jena', 'Janae', 'Gianna', 'Janay', 'Joana' |
| TN | 'Donna', 'Diana', 'Tina', 'Dena', 'Dona', 'Tennie', 'Donnie', 'Tena', 'Dagny', 'Diane', 'Dina', 'Dawn', 'Dana', 'Toni', 'Dean', 'Deanna', 'Dianne', 'Dianna', 'Diann', 'Dian', 'Dinah', 'Deanne', 'Danna', 'Tana', 'Deana', 'Deena', 'Teena', 'Deann', 'Tonia', 'Tania', 'Dionne', 'Dawna', 'Dayna', 'Tawny' |
| KL | 'Cleo', 'Callie', 'Gail', 'Gayle', 'Gale', 'Gayla', 'Kelly', 'Kelley', 'Kelli', 'Kyle', 'Kellie', 'Kayla', 'Keely', 'Kyla', 'Kylie', 'Kali', 'Kaylee', 'Kala', 'Kylee', 'Kiley', 'Kaila', 'Keila', 'Kailey', 'Kaley', 'Cayla', 'Kaylie', 'Kaela', 'Kailee', 'Kallie', 'Kayli', 'Kalie', 'Kayley' |
| KR | 'Carrie', 'Cora', 'Corrie', 'Kerry', 'Gary', 'Kari', 'Kerri', 'Cara', 'Carey', 'Kara', 'Keri', 'Cari', 'Karrie', 'Kerrie', 'Cary', 'Cori', 'Corey', 'Karie', 'Kori', 'Carie', 'Cory', 'Karri', 'Kira', 'Carri', 'Kiara', 'Kyra', 'Kiera', 'Kierra’ |
| LN | 'Lena', 'Leona', 'Leone', 'Linnie', 'Lina', 'Lona', 'Lennie', 'Lonnie', 'Luna', 'Linnea', 'Lynn', 'Lenna', 'Lonie', 'Lynne', 'Lana', 'Lyn', 'Leanna', 'Leanne', 'Luanne', 'Luana', 'Lonna', 'Luann', 'Leann', 'Louann', 'Leeann', 'Liana' |
| LS | 'Louise', 'Lucy', 'Lois', 'Lizzie', 'Lessie', 'Louisa', 'Luisa', 'Luz', 'Lucie', 'Liza', 'Loyce', 'Louis', 'Lisa', 'Lesa', 'Liz', 'Leesa', 'Lise', 'Leisa', 'Lissa', 'Lacey', 'Lacy', 'Lacie', 'Laci' |
| KRL | 'Carol', 'Coral', 'Carole', 'Creola', 'Caryl', 'Carroll', 'Carla', 'Carolee', 'Carrol', 'Carl', 'Karla', 'Karol', 'Carly', 'Carley', 'Karli', 'Karly', 'Carlie', 'Karlee', 'Karlie', 'Carli', 'Carlee', 'Karley' |
| LL | 'Lillie', 'Lula', 'Lola', 'Lela', 'Luella', 'Lila', 'Leola', 'Leila', 'Lulu', 'Lily', 'Lelia', 'Lilly', 'Louella', 'Lilla', 'Lelah', 'Lyla', 'Lilia', 'Layla' |
| KRN | 'Corinne', 'Corine', 'Corrine', 'Corene', 'Karen', 'Karin', 'Karon', 'Caren', 'Karyn', 'Karren', 'Caryn', 'Karan', 'Caron', 'Corina', 'Corinna', 'Karina', 'Carina' |
| TR | 'Dora', 'Terry', 'Terri', 'Teri', 'Terrie', 'Tara', 'Dara', 'Tori', 'Dori', 'Terra', 'Tera', 'Tyra', 'Tierra', 'Tarah', 'Tiera', 'Drew' |

In conclusion, both algorithms produce around four hundred classes of phonetic anagram classes and they both identified significant amount of similar sounding words into groups. Besides, they produce very similar top 10 largest classes which sound like “Jenny”, “Dena” and “Christine”. However it is obvious that Metaphone algorithm better identified similar sounding names, it finds those similar sounding names with even very different spelling like ‘Jenny’ and ‘Gene’. And it solves some obvious limitations in Soundex algorithm. From my examples of misclassified names, although ‘ch’ and ‘k’ sound the same in names, they start by different character, thus many commonly used name start by ‘ch’ and ‘k’ which should be placed together are misplaced into different groups, but in Metaphone algorithm, we set ‘c’ to ‘K’ which solves this issue.

**Extra credits**

25 Interesting Homonyms

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Piece** | **to** | **write** | **mail** | **by** | **sum** | **see** | **whether** | **fair** |
| **Peace** | **two** | **right** | **male** | **buy** | **some** | **sea** | **weather** | **fare** |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **know** | **plain** | **weak** | **knot** | **base** | **Sun** | **read** | **meat** |
| **no** | **plane** | **week** | **not** | **bass** | **son** | **red** | **meet** |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **mussel** | **sew** | **vary** | **for** | **which** | **here** | **feet** | **fourth** |
| **muscle** | **so** | **very** | **four** | **witch** | **hear** | **feat** | **forth** |

**Appendix**

**Code**

name\_dict = {} #create dict

with open('data.txt', 'r')as f: #import data

lines=f.readlines()

for line in lines:

name = line.strip('\n')

soundex = soundexNaive(name,len=4)

if soundex in name\_dict:

name\_dict[soundex].append(name)

else:

name\_dict[soundex] = [name]

for key in name\_dict.copy():

if len(name\_dict[key]) == 1:

del name\_dict[key]

for key in sorted(name\_dict, key=lambda key: len(name\_dict[key]), reverse=True):

print(key,name\_dict[key])

I use the same structure from the anagrams program for finding the name phonetic anagram classes. So for each of Soundex and Metaphone output, I build up a dictionary in python. I set the key to be the Soundex encoding/Metaphone encoding and value is all the names have that code. An example of Soundex process diagram is shown below:

**Reference**

[1] <https://en.wikipedia.org/wiki/Soundex>

[2] <http://www.blackwasp.co.uk/soundex.aspx>

[3] <https://en.wikipedia.org/wiki/Phonetic_algorithm>

[4] <https://en.wikipedia.org/wiki/Phonetic_algorithm>