به نام خدا

تمرین شماره 1

1. تلفظ عبارات فارسی، وابسته به کلماتی که قبل یا بعد از هر کلمه آورده می شوند ممکن است ابهام برانگیز باشد. هدف از این تمرین پیدا کردن ابهام در گفتار جملات فارسی است. حاصل کار شما باید یک عبارت با جمله فارسی را از ورودی بگیرد و لیست تمامی جملات (یا عبارات) غیر یکسانی را در خروجی قرار دهد که معادل گفتار معادل هستند. به عنوان مثال:

ورودی:

[باغبانان] [باغچه] [را] [نگاه] [کرد]
[باغبانان] [باغ] [چه] [را] [نگاه] [کرد]
[باغبانان] [باغ] [چرا] [نگاه] [کرد]
«باغبان آن باغچه را نگاه کرد. »

خروجی:

[باغبان] [آن] [باغ] [چه] [را] [نگاه] [کرد]
[باغبان] [آن] [باغ] [چرا] [نگاه] [کرد]
[باغ] [با] [نان] [باغچه] [را] [نگاه] [کرد]
[باغ] [با] [نان] [باغ] [چه] [را] [نگاه] [کرد]
[باغ] [با] [نان] [باغ] [چرا] [نگاه] [کرد]

در واقع جمله ی «باغبان آن باغچه را نگاه کرد. » با توجه به دیکشنری های موجود دارای فونتیکی به صورت زیر است:

و تمامی جملات خروجی نیز دارای همین مجموعه ی به هم پیوسته ی bAqbAn+An+bAqCe+rA+negAh+kard. فونتیک هستند

برای دسترسی آسان تر دیکشنری ای حاوی فونتیک کلمات فارسی در اختیار شما خواهد بود. این دیکشنری به صورت یک فایل با فرمت csv خواهد بود که در دو ستون WrittenForm و PhonologicalForm به ترتیب شکل نوشتاری و شکل گفتاری هر کلمه را خواهید یافت.

2. با استفاده از هر زبان برنامه نویسی برنامه ای با استفاده از regex بنویسید که تمامی کلمات زیر را پوشش دهد

- William R. Breakey M.D.
- Pamela J. Fischer M.D.
- Leighton E. Cluff M.D.
- James S. Thompson, M.D.
- C.M. Franklin, M.D.
- Atul Gawande, M.D.
- Dr. Talcott
- Dr. J. Gordon Melton
- Dr. Etienne-Emile Baulieu
- Dr. Karl Thomae
- Dr. Alan D. Lourie
- Dr. Xiaotong Fei
- Doctor Dre
- Doctor Dolittle
- Doctor William Archibald Spooner
- Doctor No

برای تست عملکرد برنامتون این متن ها رو باید توی یه فایل تکست بریزید و ازش بخونید و پیدا کردنش رو با 0 یا 1 نمایش بدید.

3. برنامه ای بنویسید که یک فایل شامل دو عبارت از ورودی بخواند و Minimum Edit Distance دو عبارت رو خروجی بدهد

4. (امتیازی) spell correction امروزه کاربرد زیادی در سیستم های کامپیوتری دارد یکی از راه های مقدماتی پیاده سازی آن با استفاده از MED است
در این تمرین شما باید یک spell correction سیستم پیاده سازی کنید که با گرفتن یک کلمه به عنوان ورودی 3 کلمه که کمترین MED را با آن دارند در خروجی نمایش بدهد
   - برای دیتاست کلمات انگلیسی می توانید از این لینک استفاده کنید
   - بدیهی است هر چقدر که برنامه شما عملکرد بهتری داشته باشد نمره بیشتری دریافت خواهید کرد
   - برای کم شدن فضای جستجو می توانید فرض کنید که کلمه اول عبارت همیشه درست نوشته شده است.

برای تمامی سوالات یک پوشه به اسم src که شامل کد هاتون باشه و یک پوشه به اسم in که فایل ورودی داخل اون باشه و از اونجا خونده بشه ورودی همچنین یک پوشه out که خروجی ها و نتایج داخل اون باشند وجود داشته باشه و یه فایل readme هم توی پوشه هر سوال باشه که توش روش حل مساله رو توضیح داده باشید(الزامی)
مثلا

- Problem1
  - readme.md
  - src
    - med.py
  - out
    - result.txt
  - in
    - input.txt