# CSE 654/484 HW2

Can not seperate words into  syllables so I used words in n-grams instead of syllables.

Tables:

For 1-gram,

| Words | Counts |
|---|---|
| ('<doc',) | 4 |
| ('id="10"',) | 1 |
| ('url="https://tr.wikipedia.org/wiki?curid=10"',) | 1 |
| ('title="cengiz',) | 1 |
| ('han">',) | 1 |
| ('cengiz',) | 309 |
| ('han',) | 131 |
| ('("cenghis',) | 1 |
| ('khan",',) | 1 |
| ('"cinggis',) | 1 |
| ('haan"',) | 1 |
| ('ya',) | 10 |
| ('da',) | 89 |
| ('dogum',) | 7 |
| ('adiyla',) | 2 |
| ('temucin',) | 35 |
| ('(anlami:',) | 2 |
| ('hukumdar',) | 7 |
| ('ve',) | 436 |
| ("İmparatorlugu'nun",) | 1 |
| ('kurucusudur.',) | 1 |
| ('han,',) | 59 |
| ('13.',) | 3 |
| ('yuzyilin',) | 2 |
| ('basinda',) | 7 |
| ('orta',) | 10 |
| ("asya'daki",) | 2 |
| ('tum',) | 16 |
| ('gocebe',) | 5 |
| ('bozkir',) | 2 |
| ('kavimlerini',) | 1 |
| ('birlestirerek',) | 2 |
| ('bir',) | 334 |
| ('ulus',) | 3 |
| ('haline',) | 6 |
| ('getirdi',) | 2 |
| ('o',) | 17 |

For 2-gram,

| Words | Counts |
|---|---|
| ('<doc', 'id="10"') | 1 |
| ('id="10"', 'url="https://tr.wikipedia.org/wiki?curid=10"') | 1 |
| ('url="https://tr.wikipedia.org/wiki?curid=10"', 'title="cengiz') | 1 |
| ('title="cengiz', 'han">') | 1 |
| ('han">', 'cengiz') | 1 |
| ('cengiz', 'han') | 113 |
| ('han', 'cengiz') | 1 |
| ('han', '("cenghis') | 1 |
| ('("cenghis', 'khan",') | 1 |
| ('khan",', '"cinggis') | 1 |
| ('"cinggis', 'haan"') | 1 |
| ('haan"', 'ya') | 1 |
| ('ya', 'da') | 10 |
| ('da', 'dogum') | 1 |
| ('dogum', 'adiyla') | 1 |
| ('adiyla', 'temucin') | 1 |
| ('temucin', '(anlami:') | 1 |

| | |
|---|---|
| ('catisi', 'altinda') | 2 |
| ('altinda', 'topladi.') | 1 |
| ('topladi.', 'dunya') | 1 |
| ('dunya', 'tarihinin') | 1 |
| ('tarihinin', 'en') | 3 |
| ('en', 'buyuk') | 17 |
| ('buyuk', 'askeri') | 1 |
| ('askeri', 'dehalarindan') | 1 |
| ('dehalarindan', 'biri') | 1 |
| ('biri', 'olarak') | 6 |
| ('olarak', 'kabul') | 6 |
| ('kabul', 'edilen') | 2 |
| ('edilen', 'cengiz') | 3 |
| ('han,', 'hukumdarligi') | 1 |
| ('hukumdarligi', 'doneminde') | 1 |
| ('doneminde', '1206-1227') | 1 |
| ('1206-1227', 'arasinda') | 1 |
| ('arasinda', 'kuzey') | 1 |
| ('kuzey', "cin'deki") | 3 |

For 3-gram,

| Words | Counts |
|---|---|
| ('<doc', 'id="10"', 'url="https://tr.wikipedia.org/wiki?curid=10"') | 1 |
| ('id="10"', 'url="https://tr.wikipedia.org/wiki?curid=10"', 'title="cengiz') | 1 |
| ('url="https://tr.wikipedia.org/wiki?curid=10"', 'title="cengiz', 'han">') | 1 |
| ('title="cengiz', 'han">', 'cengiz') | 1 |
| ('han">', 'cengiz', 'han') | 1 |
| ('cengiz', 'han', 'cengiz') | 1 |
| ('han', 'cengiz', 'han') | 1 |
| ('cengiz', 'han', '("cenghis') | 1 |
| ('han', '("cenghis', 'khan",') | 1 |
| ('("cenghis', 'khan",', '"cinggis') | 1 |
| ('khan",', '"cinggis', 'haan"') | 1 |
| ('"cinggis', 'haan"', 'ya') | 1 |
| ('haan"', 'ya', 'da') | 1 |
| ('ya', 'da', 'dogum') | 1 |
| ('da', 'dogum', 'adiyla') | 1 |
| ('dogum', 'adiyla', 'temucin') | 1 |
| ('adiyla', 'temucin', '(anlami:') | 1 |
| ... | ... |
| ('cengiz', 'han', 'soyunun') | 2 |
| ('han', 'soyunun', 'efsanevi') | 1 |
| ('soyunun', 'efsanevi', 'buyuk') | 1 |
| ('efsanevi', 'buyuk', 'annesi') | 1 |
| ('buyuk', 'annesi', 'olarak') | 1 |
| ('annesi', 'olarak', 'kabul') | 1 |
| ('olarak', 'kabul', 'edilmistir.') | 2 |
| ('kabul', 'edilmistir.', 'mogollarin') | 1 |
| ('edilmistir.', 'mogollarin', 'gizli') | 1 |
| ('mogollarin', 'gizli', 'tarihinde') | 2 |
| ('gizli', 'tarihinde', 'yer') | 1 |
| ('tarihinde', 'yer', 'alan') | 1 |
| ('yer', 'alan', 'efsaneye') | 1 |
| ('alan', 'efsaneye', 'gore') | 1 |
| ('efsaneye', 'gore', 'alangoya') | 1 |
| ('gore', 'alangoya', 'dul') | 1 |
| ('alangoya', 'dul', 'kaldiktan') | 1 |
| ('dul', 'kaldiktan', 'sonra') | 1 |
| ('kaldiktan', 'sonra', 'evlenmedigi') | 1 |
| ('sonra', 'evlenmedigi', 'halde') | 1 |

Good Turing Smoothing for all of them,

For 1-gram,

```
{('<doc',): 0.00040429710061221384,
 ('id="10"',): 4.4822540261693854e-05,
 ('url="https://tr.wikipedia.org/wiki?curid=10"',): 4.4822540261693854e-05,
 ('title="cengiz',): 4.4822540261693854e-05,
 ('han">',): 4.4822540261693854e-05,
 ('cengiz',): 0.0,
 ('han',): 0.0,
 ('("cenghis',): 4.4822540261693854e-05,
 ('khan",',): 4.4822540261693854e-05,
 ('"cinggis',): 4.4822540261693854e-05,
 ('haan"',): 4.4822540261693854e-05,
 ('ya',): 0.000782719186785246,
 ('da',): 0.0,
 ('dogum',): 0.000992878461038367,
 ('adiyla',): 0.00014057621875844206,
 ('temucin',): 0.0,
 ('(anlami:',): 0.00014057621875844206,
 ('demirci),',): 4.4822540261693854e-05,
```

For 2-gram,

```
{('<doc', 'id="10"'): 4.194027590535075e-05,
 ('id="10"',
  'url="https://tr.wikipedia.org/wiki?curid=10"'): 4.194027590535075e-05,
 ('url="https://tr.wikipedia.org/wiki?curid=10"',
  'title="cengiz'): 4.194027590535075e-05,
 ('title="cengiz', 'han">'): 4.194027590535075e-05,
 ('han">', 'cengiz'): 4.194027590535075e-05,
 ('cengiz', 'han'): 0.0,
 ('han', 'cengiz'): 4.194027590535075e-05,
 ('han', '("cenghis'): 4.194027590535075e-05,
 ('("cenghis', 'khan",'): 4.194027590535075e-05,
 ('khan",', '"cinggis'): 4.194027590535075e-05,
 ('"cinggis', 'haan"'): 4.194027590535075e-05,
 ('haan"', 'ya'): 4.194027590535075e-05,
 ('ya', 'da'): 0.0054536440257795344,
 ('da', 'dogum'): 4.194027590535075e-05,
 ('dogum', 'adiyla'): 4.194027590535075e-05,
```

For 3-gram,

```
('teskilati', 'kullanarak', 'meritokratik'): 4.8930175458529156e-05,
('kullanarak', 'meritokratik', '(liyâkata'): 4.8930175458529156e-05,
('meritokratik', '(liyâkata', 'bagli)'): 4.8930175458529156e-05,
('(liyâkata', 'bagli)', 'bir'): 4.8930175458529156e-05,
('bagli)', 'bir', 'ordu'): 4.8930175458529156e-05,
('bir', 'ordu', 'meydana'): 0.0012507600104225323,
('ordu', 'meydana', 'getiren'): 4.8930175458529156e-05,
('meydana', 'getiren', 'cengiz'): 4.8930175458529156e-05,
('getiren', 'cengiz', 'han’in'): 4.8930175458529156e-05,
('cengiz', 'han’in', 'buyuk'): 0.02015113350125191,
('han’in', 'buyuk', 'bir'): 0.0012507600104225323,
('buyuk', 'bir', 'asker'): 0.0012594458438282444,
('bir', 'asker', 'olarak'): 0.0012594458438282444,
('asker', 'olarak', 'un'): 0.0012594458438282444,
('olarak', 'un', 'kazanmasinin'): 0.0012594458438282444,
('un', 'kazanmasinin', 'temelinde,'): 4.8930175458529156e-05,
('kazanmasinin', 'temelinde,', 'kurdugu'): 4.8930175458529156e-05,
('temelinde,', 'kurdugu', 'posta'): 4.8930175458529156e-05,
('kurdugu', 'posta', 'teskilati'): 0.0012507600104225323,
('posta', 'teskilati', 've'): 0.0012507600104225323
```

Can not complete rest of the homework

Muhammed Alperen Karaçete 171044052