Ders Adı : Doğal Dil İşlemeye Giriş
Yürütücü İsmi : Prof. Dr. Banu Diri
Proje Adı : Sovyetler Birliği Dilleri Yapısal Benzerlik Analizi
Proje Teslim Tarihi : 03.01.2023
Öğrenci Ad ve Soyadları :

16011702 - Mustafa Aydın
17011615 - Said Eroğlu

1. Proje Konusu

Proje, eski Sovyetler Birliği ülkesi dillerinin benzerliklerini incelemeyi amaçlamıştır. Dil, bilim ve sanat için oldukça önemlidir. Dünyadaki her bir dil, zenginlikleri açısından benzersizdir. Tarih boyunca da diller, zamanla birbirlerinden etkilenmişlerdir. Bir dönem aynı çatı altında bulunan Kafkas, Türk, Slav ve Baltık milletlerinin bu durumda bulunmaları kaçınılmaz olmuştur. Dünyanın birçok yerinde, bir kasabadan diğerine küçük dil farklılıkları, yarım saatlik bir yolculukta tamamen farklı bir dil oluşturur. Nefes kesen karmaşıklığı ve çeşitliliği için insan dili, zaman ve mekan boyunca uzanan renkli bir duvar halısı gibidir.

2. Geliştirilen Sistem Blok Şeması

2.1. NLP Pipeline

String veri türü olan dizelerle ilgilenmek için, yerel dosya içeriğini yükleriz. Ardından dize, simgeleştirilir ve kelime listesi üretilir. Listeler önce normalleştirilip sonra sıralanır.

```
tokens = nltk.word_tokenize(raw)
words = [w.lower() for w in tokens]
vocab = sorted(set(words))
```

2.2. Prosedürel ve Bildirime Dayalı Stil

Programlama stili, program geliştirmeyi etkiler. Yer yer kodda, bilgisayar CPU'su tarafından gerçekleştirilen ilkel işlemler olan makine kodundan çok uzak olmayan bir stil benimsenmiştir. Program, prosedürler ve makine işlemlerini adım adım dikte eder.

```
count = 0
total = 0
for token in tokens:
    count += 1
    total += len(token)
print("Ortalama kelime uzunluğu : %2.2f" % (total/count))
```

Her çıktı satırıyla bir sayaç yazdırmak için gerekli görüldüğünde döngü sayacı kullanılmıştır. Aşağıdaki kod, etiket seti oluşturmak adına yazılmıştır. Frekans dağılımı anahtarları sıralanır ve word değişkenlerinde tamsayı-string çifti yakalanır. Sıralı öğelerin bir listesi üretildiğinde, sayım 1'den başlıyor gibi görünecek şekilde rank + 1 yazdırılır.

```
cumulative = 0.0
tagged_tokens = [[]]
for rank, word in enumerate(fdist1): # %25'e kadar yer kaplayan kelimeler
    cumulative += fdist1[word] * 100 / fdist1.N()
    print("%3d %6.2f%% %s" % (rank+1, cumulative, word))
    tagged_tokens[0].append((word, word))
    if cumulative > 25:
        break
```

Programda istenen şeylerden biri de, en uzun kelimelerin tümünü bulmak.

```
longest = ''
for word in vocab:
    if len(word) > len(longest):
        longest = word
print("\nEn uzun kelimeler : ")
maxlen = max(len(word) for word in vocab) # En büyük uzunluk
print([word for word in vocab if len(word) == maxlen])
```

2.3. Frekans Dağlımları

Sıklık dağılımları, bize metindeki her kelime öğesinin sıklığını söyler; genel olarak, her türlü gözlemlenebilir olayın. Toplam kelime belirteçlerinin, kelime üyeleri arasında nasıl dağıldığını bize anlatır.

İlgili metin kelimelerinin frekans dağılımı :

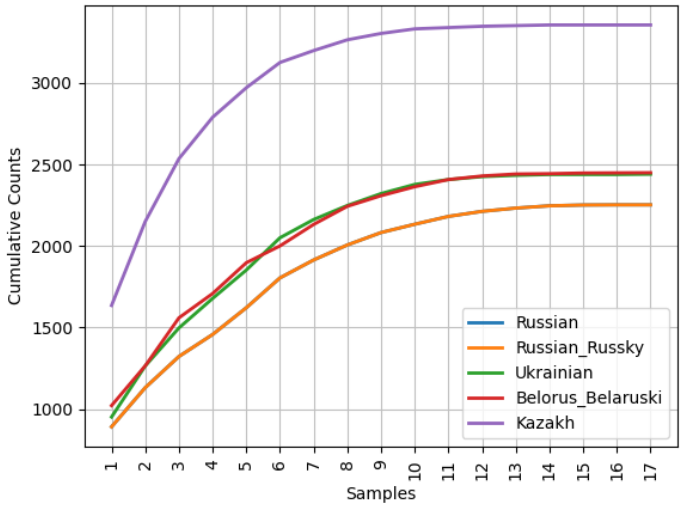
```
tokens = nltk.word_tokenize(raw)
words = [w.lower() for w in tokens]
vocab = sorted(set(words))
fdist1 = FreqDist(words)
```

Hedef metnin, etiketlenmiş kelime bazında frekans dağılımı :

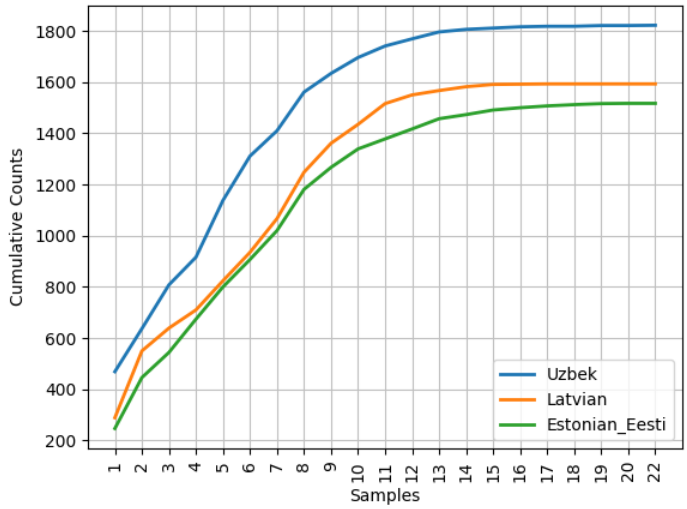
```
s_tokens = nltk.word_tokenize(raw2)
unigram_tagger = nltk.UnigramTagger(tagged_tokens)
tagged_sentence = unigram_tagger.tag(s_tokens)
fdist2 = FreqDist(tagged_sentence)
```

2.4. Frekans Dağılımlarını Çizme ve Tablolama

Projemizde, dağılımlarımızın koşulu dildir. Ayrıca projeye ek olarak, İnsan Hakları Evrensel Beyannamesi örnek alınarak dillerin kelime uzunluk graf ve tabloları çizdirildi.



Şekil 1. Kümülatif kelime uzunluğu dağılımları: İnsan Hakları Evrensel Beyannamesi'nin dört çevirisi işlenir (Kiril alfabesi);



Şekil 2. Kümülatif kelime uzunluğu dağılımları: İnsan Hakları Evrensel Beyannamesi'nin üç çevirisi işlenir (Latin alfabesi);

				(KİRİL)					
	0	1	2	3	4	5	6	7	8	9
Russian_Russky	0	892	1132	1323	1458	1622	1805	1915	2007	2083
Ukrainian	0	952	1265	1498	1679	1853	2051	2163	2249	2322
Belorus_Belaruski	0	1022	1266	1560	1709	1898	2001	2133	2244	2309
Kazakh	0	1636	2151	2536	2790	2971	3126	3199	3265	3304

Tablo 1. *Bu tablo, beş veya daha az harfli kelimelerin Rusça, Uktaynaca ve Belarusça metinlerinin yaklaşık %80'ini ve Kazakça metninin %90'ını oluşturduğunu göstermektedir.*

				(LATİN)					
	0	1	2	3	4	5	6	7	8	9
Uzbek	0	469	637	808	916	1138	1311	1410	1561	1634
Latvian	0	289	550	639	710	824	935	1067	1248	1361
Estonian_Eesti	0	247	446	544	674	800	907	1020	1181	1267

Tablo 2. *Bu tablo, beş veya daha az harfli kelimelerin Özbekçe metninin yaklaşık %70'ini ve Letonca ve Estonca metinlerinin %60'ını oluşturduğunu göstermektedir.*

2.6. Metin Kaydırma

Program çıktısı tablo yerine metin benzeri olduğunda, uygun şekilde görüntülenebilmesi için sarıldı. Yazım formatı olarak da "C dili" sözdizimi benimsendi.

```
print("\nBenzerlik yüzdesi(Estonca) : %2.2f%%" % percentage)
```

3. Veri Seti

3.1. Corpora'lar

Dosya kimlikleri, Latin ve Kiril kodlamasına sahiptir. Wikipedia külliyatında yer alan çeşitli diller kullanıldı.

Kiril Veriler :

Dil	Kaynak	Yıl	Cümle Sayısı	Boyut
Belarusça	Wikipedia	2016	30 K	5.8 MB
Kazakça	Wikipedia	2014	30 K	6.4 MB
Kırgızca	Wikipedia	2016	30 K	5.7 MB
Rusça	Wikipedia	2014	30 K	6.0 MB
Tacikçe	Wikipedia	2016	30 K	5.7 MB
Uktaynaca	Wikipedia	2016	30 K	6.1 MB

Latin Veriler :

Dil	Kaynak	Yıl	Cümle Sayısı	Boyut
Azerice	Wikipedia	2016	30 K	3.9 MB
Estonca	Wikipedia	2014	30 K	3.5 MB
Letonca	Wikipedia	2014	30 K	3.5 MB
Litvanca	Wikipedia	2014	30 K	3.2 MB
Rumence	Wikipedia	2014	30 K	4.1 MB
Türkmence	Wikipedia	2016	30 K	3.5 MB
Özbekçe	Wikipedia	2014	30 K	3.4 MB

Ne yazık ki, birçok dil için önemli bir derlem henüz mevcut değil. Dil kaynaklarını geliştirmek için genellikle hükümet veya endüstriyel destek yetersizdir ve bireysel çabalar parça parçadır ve keşfedilmesi veya yeniden kullanılması zordur. Bazı dillerin yerleşik bir yazı sistemi yoktur veya tehlike altındadır.

3.2. Yerel Dosyada Okuma

Yerel bir dosyayı okumak için Python'un yerleşik open() ve read() işlevleri kullanıldı.

```
f = open('Sentences/Latin/aze_wikipedia_2016_30K-sentences.txt')
raw = f.read()
```

3.3. Veri Seti Kaynakları

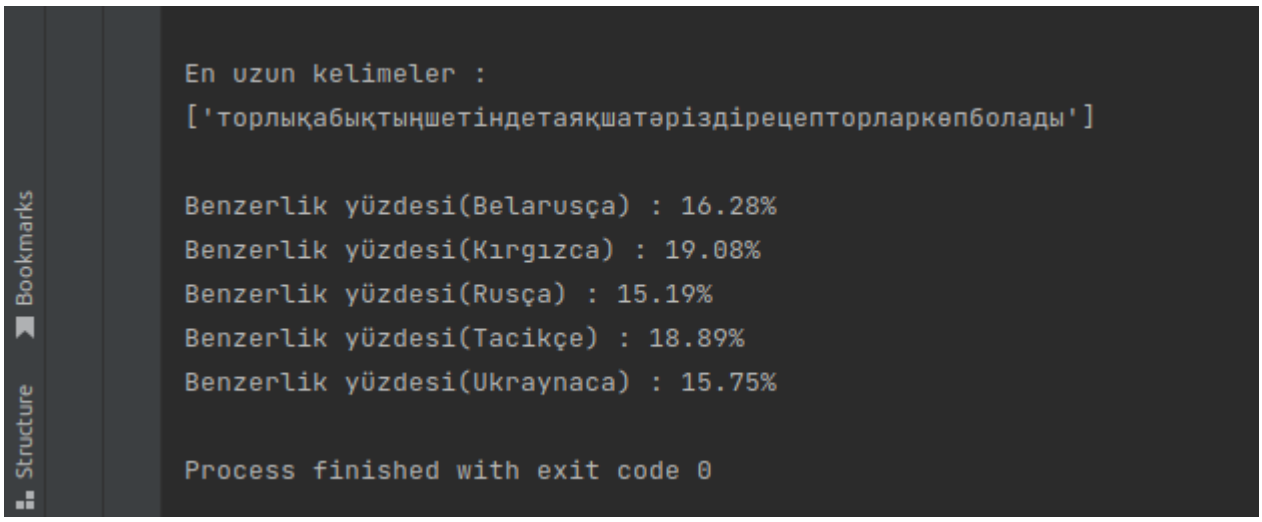
Leipzig Corpora Collection / Deutscher Wortschatz
https://wortschatz.uni-leipzig.de/en

Kullanılan metin materyali,
- Çeşitli kaynaklardan, (haber materyali, Web metni vb.)
- CURL portalı (https://curl.corpora.uni-leipzig.de) aracılığıyla toplanan metin materyalinden,
- Haber sitelerinden, (genellikle RSS beslemeleri aracılığıyla günlük olarak)
- Haber sitelerinden, (belirtilen yıldan daha eski olabilir.)
- Rastgele seçilmiş Web sitelerinden,
- Wikipedia dökümlerinden

alınmıştır.

4. Örnek Program Çıktıları

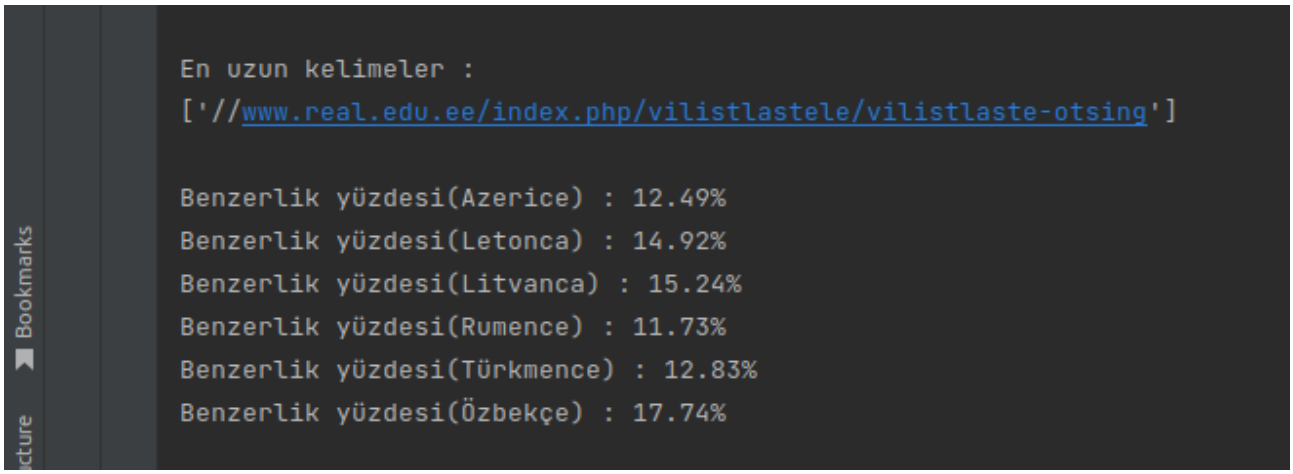
4.1. Kazakça



Oluşan etiket seti :

1 5.73% .	2 10.91% ,
3 12.24% және	4 13.22% мен
5 13.98%)	6 14.73% (
7 15.46% бұл	8 15.96% оның
9 16.42% да	10 16.87% –
11 17.32% болып	12 17.77% бар
13 18.21% де	14 18.62% бір
15 19.03% »	16 19.43% «
17 19.77% үшін	18 20.10% :
19 20.41% —	20 20.73% ол
21 21.03% деп	22 21.33% бойынша
23 21.61% немесе	24 21.89% жылы
25 22.17% деген	26 22.44% -
27 22.69% осы	28 22.93% болды
29 23.15% болады	30 23.35% сондай-ақ
31 23.56% болатын	32 23.76% екі
33 23.96% болған	34 24.14% сол
35 24.32% арқылы	36 24.50% басқа
37 24.67% байланысты	38 24.84% адам
39 25.01% өз	

4.2. Estonca



Oluşan etiket seti :

1 5.96% .	2 10.11% ,
3 12.50% ja	4 14.45% on
5 15.63% (	6 16.81%)
7 17.78% oli	8 18.51% ta
9 19.07% ning	10 19.60% eesti
11 20.12% ka	12 20.55% ``
13 20.97% "	14 21.36% kui
15 21.75% –	16 22.14% et
17 22.52% ei	18 22.84% aastal
19 23.17% mis	20 23.43% see
21 23.67% aasta	22 23.91% :
23 24.15% oma	24 24.35% kuid
25 24.54% tema	26 24.72% sai
27 24.90% või	28 25.07% selle

5. Total Benzerlik Tablosu

Sheet1

	Belarusça	Kazakça	Kırgızca	Rusça	Tacikçe	Ukraynaca
Belarusça		13.76	16.85	16.31	16.66	19.82
Kazakça	16.28		19.08	15.19	18.89	15.75
Kırgızca	16.27	15.99		15.31	18.82	15.68
Rusça	16.89	13.3	16.5		16.68	17.85
Tacikçe	13.49	12.47	15.38	12.57		12.79
Ukraynaca	21.49	13.57	17.11	19.43	17.74	

	Azerice	Estonca	Letonca	Litvanca	Rumence	Türkmence	Özbekçe
Azerice		13.62	14.75	15.31	12.71	14.35	18.12
Estonca	12.49		14.92	15.24	11.73	12.83	17.74
Letonca	12.14	13.35		16.84	12.05	12.36	17.22
Litvanca	11.71	13.17	15.55		12.28	13.14	16.87
Rumence	9.85	10.16	11.2	12.09		10.96	12.21
Türkmence	13.22	13.17	13.83	15.84	11.49		17.21
Özbekçe	12.96	13.33	14.7	14.29	11.42	12.79	

Birbirlerine en çok benzeyen diller :

1. Belarusça ve Ukraynaca
2. Kazakça ve Kırgızca
3. Azerice ve Özbekçe