בס"ד

UNIGRAMNS – ANSWERS

Unigrams (10 points)

You should produce answers to the following questions. You probably want to do it by writing one or several short programs. Python would be a good choice, but we leave the decision up to you.

For each of the files (Hebrew, English):

1. How many tokens are in the file?
2. How many distinct types are in the file?
3. What is the tokens/types (tokens divided by types) ratio for each language? Can you explain this difference?
4. Now, let’s look at the first half of each file (if the number of tokens in the file is N, we look at the first N/2 tokens). How many distinct types are in this first half? How many distinct types were added when going from the first half to the entire file? Are there types that don’t appear in the first half, but appear more than two times in the second half? more than 5 times? more than 10 times? provide examples of each, or say that none exist.
5. Back to the full files: what are the top 50 most frequent token-types in each file (sorted from most frequent to least frequent)? Are the lists similar or different? Briefly discuss the reasons for this.
6. Which is the 100th most frequent word in each language? What is the 1,000th most frequent? What is the 10,000 most frequent?
7. In each language, how many types appear: exactly once? exactly twice? exactly three times? exactly 5 times? how many types appear 10 times or more? 100 times or more? 1,000 times or more? 10,000 times or more? Arrange these numbers in a table.

**Analysis for English:**

**1. Total tokens:** 84512402

2. **Distinct types in full file**: 868910

**3. Tokens/Types:** 97.26254963114707

4**. Distinct types in first half**: 582710

5**. Types added in second half:** 286200

6**. Examples of types appearing only in second half:**

> More than 2 times: ['Wigley', 'boater', "Quint's", "Scheider's", "Sackler's"]

> More than 5 times: ['boater', "Quint's", 'Mattey', 'Heman', 'carny']

> More than 10 times: ["Quint's", 'Heman', 'head-tefillin', 'Ashkenazim', 'Shawnees']

**Top 50 tokens**:

the: 5581841

of: 2733726

and: 2494724

to: 1973589

in: 1884049

a: 1674060

was: 1075262

The: 853610

on: 683233

that: 676753

as: 674713

for: 674483

with: 631925

by: 598630

is: 512726

his: 436296

from: 431770

at: 415899

were: 354530

it: 318207

he: 315354

an: 305976

had: 292521

In: 281283

which: 275589

be: 234529

but: 203280

are: 203194

not: 194652

their: 186417

first: 182481

also: 179718

her: 176737

its: 174838

or: 166247

have: 163406

who: 154883

been: 152075

one: 151840

has: 150710

this: 145553

two: 144315

He: 139795

they: 133512

after: 129988

would: 126043

into: 123404

time: 115312

other: 113333

It: 112011

**Frequency Buckets:**

**10+:** 153429

**100+**: 38442

**1,000**+: 7919

**10,000+**: 903

**2**: 121095

**1**: 418423

**3:** 58461

**5**: 25017

**Analysis for Hebrew:**

1**. Total tokens:** 63556314

**2. Distinct types in full file:** 1627654

**3. Tokens/Types:** 39.047803771563245

**4. Distinct types in first half:** 1022933

5**. Types added in second half:** 604721

**6. Examples of types appearing only in second half:**

> More than 2 times: ['המרסלה', 'למאפייניהם', 'Riserva', 'Soleras', 'והבובנאי']

> More than 5 times: ['הטיטניט', 'לבראונינג', 'Heelers', 'אלטש', 'ביטשהורן']

> More than 10 times: ['אלטש', 'ביזי', 'DAF-2', 'הבילס', 'הווילד']

**Top 50 tokens**:

של: 1689076

את: 976603

על: 838176

הוא: 439643

עם: 322381

בשנת: 232544

היה: 226924

גם: 215117

היא: 205409

לא: 203901

לאחר: 192311

ידי: 187061

בין: 186277

או: 177070

זה: 152923

כי: 149337

אך: 132133

עד: 123229

יותר: 123200

כל: 120793

זו: 112623

אשר: 110476

הייתה: 108598

ישראל: 101607

הם: 95060

ביותר: 92762

היו: 85651

בית: 85475

הראשון: 85295

כמו: 83782

כאשר: 82841

אחד: 82596

שם: 81932

בו: 81233

מספר: 80409

הברית: 77503

כך: 74400

כדי: 70647

אותו: 69497

שלו: 68735

יש: 68468

אל: 68002

העיר: 66140

במהלך: 65476

לו: 65301

שני: 64956

רבים: 62832

זאת: 61247

אף: 61245

רק: 60859

**Frequency Buckets:**

1**:** 881206

**10**+: 241655

**100**+: 54650

**1,000**+: 8740

**10**,**000**+: 641

**2**: 212129

**3**: 99628

**5**: 41609

Process finished with exit code 0

English :

Token counts written to tokens\_english.txt.txt

1 - tokens count: 85670540

2 - distinct tokens count: 769201

3 - tokens\_counts / distinct tokens count : 111.37601225167414

Hebrew :

Token counts written to tokens\_hebrew.txt.txt

1 - tokens count: 65369174

2 - distinct tokens count: 1375351

3 - tokens\_counts / distinct tokens count : 47.52908457550109

Analysis for English:

1. Total tokens: 85670540

2. Distinct types in first half: 518783

3. Distinct types in full file: 769201

4. Types added in second half: 250418

5. Examples of types appearing only in second half:

> More than 2 times: ['Duberry', 'Ancion', 'Wallwork', 'Wigley', 'Rolly']

> More than 5 times: ['Wallwork', 'Mattey', 'Heman', 'carny', 'X-23']

> More than 10 times: ['Wallwork', 'Heman', 'head-tefillin', 'Ashkenazim', '7200']

Top 50 tokens:

the: 5581945

of: 2734013

and: 2494918

to: 1973730

in: 1884204

a: 1676811

was: 1075307

The: 853790

s: 747266

on: 683363

that: 679678

as: 674731

for: 674493

with: 631955

by: 598633

is: 512752

his: 436345

from: 431809

at: 415986

were: 354532

it: 326264

he: 318285

an: 306463

had: 292526

In: 281327

which: 275585

be: 234539

but: 203280

are: 203195

not: 194650

their: 186424

first: 182429

also: 179718

her: 176741

its: 174844

or: 166284

have: 163405

who: 155401

one: 152987

been: 152076

has: 150709

this: 145550

two: 144273

He: 141122

they: 135538

after: 129991

would: 126137

into: 123404

It: 119195

time: 115314

Frequency Buckets:

10+: 142749

100+: 37933

1,000+: 7898

10,000+: 922

2: 106634

1: 363880

3: 51599

5: 22110

Analysis for Hebrew:

1. Total tokens: 65369174

2. Distinct types in first half: 887700

3. Distinct types in full file: 1375351

4. Types added in second half: 487651

5. Examples of types appearing only in second half:

> More than 2 times: ['מגאטון', 'סאראבהאי', 'הגיאופיזית', 'Insulin-like', 'DAF-2']

> More than 5 times: ['סאראבהאי', 'DAF-2', 'Gilman', 'באושוויינצ', 'בנדט']

> More than 10 times: ['DAF-2', 'הבילס', 'הווילד', 'איינסלי', 'ורקילדסן']

Top 50 tokens:

של: 1689417

את: 977225

על: 839433

הוא: 439955

עם: 322539

-: 293055

בשנת: 232671

היה: 227001

גם: 215211

היא: 205607

לא: 204940

לאחר: 192515

ידי: 187357

בין: 186447

או: 183871

זה: 153159

ג: 151041

כי: 150577

אך: 132297

עד: 123351

יותר: 123229

כל: 121164

זו: 113091

אשר: 110540

הייתה: 108621

ישראל: 101897

ל: 99404

הם: 95196

ביותר: 92755

בית: 88822

היו: 85978

הראשון: 85309

כמו: 83992

כאשר: 82907

אחד: 82747

ה: 82589

בו: 82199

שם: 82119

ב: 81045

מספר: 80498

הברית: 77634

כך: 74694

ר: 74489

כדי: 70672

אל: 69914

יש: 69706

אותו: 69540

לו: 69540

שלו: 68802

העיר: 66220

Frequency Buckets:

10+: 235770

100+: 54735

1,000+: 8980

10,000+: 676

1: 689708

2: 183154

3: 88588

5: 38276

Process finished with exit code 0