Report for HW3

Quick Start

To run everything, simply place

- All default files from `hw3.tar.gz`. Please unzip `corpus.de.gz` and `corpus.en.gz` as well. The directory should contain

```
corpus.de
corpus.en
alignment_sample_model1.txt
alignment_sample_model2.txt
eval_scramble.py
sample_t_model1.txt
original.de
scrambled.en
devwords.txt
original.en
```

- `run.sh`
- Java files

```
Counts.java
IBM1.java
IBM2.java
Test.java
Unscramble.java
```

into the same folder.

Then run

```
sh run.sh
```

To run all of question 4, 5, and 6, and see outputs and run time.

The program allocates **1024m** of memory as minimum, with maximum of **2048m** to allow for the EM algorithm to run efficiently using `java -Xms1024m -Xmx2048m`. This is not unreasonable.

Program Structure

- `IBM1.java` represents the IBM1 model. It uses `Counts.java` to get and put counts for c(e|f), c(e), c(j|ilm), c(ilm).
- `IBM2.java` represents the IBM2 model. It extends `IBM1.java`
- `Counts.java` keeps track of the 4 counts mentioned, simplying get and put operations. For c(j|ilm) and c(ilm), it hashes ilm into a single number, simplifying and speed up retrieval.
- `Unscramble.java` deals with part 6.

Both `IBM1.java` and `IBM2.java` can have their t-table and q-table serialized onto the harddisk for quicker testing. However, `run.sh` runs the program in its full glory.

`Test.java` is the test procedure we run for the sake of this homework. It does

1. Runs IBM1 model by reading in corpuses, initializing t values, and running EM on the t values.
2. Passes IBM1's t tables to IBM2.
3. IBM2 initializes q values, runs EM on both the t values and the q values.
4. IBM2's object gets passed to Unscramble, which uses the t table and q table to unscramble sentences.

This is only one of the possible ways to use this code. Too lazy to document the rest since this fulfills the homework.

Runtime

Runtime is fast for EM due to data structure optimization. For example, I wrapped the counts in a Counts class that is built on hashtables, and hashed `i, l, m` such that the three integers can be hashed to a single number.

All operations for question 4 5 and 6 takes around 6 minutes on my computer.

Results

Question 4

The data structure for this question is optimized using hashtables and a Count class.

4.1

No answer required.

4.2

Answers are recorded to `ibm1_devwords_ranking.txt`. Reproduced here.

```
i
[('ich', 0.6981572060006594), (',', 0.07326584837825492), ('.', 0.036776505426203907),

dog
[('delendum', 0.07261346244657423), ('stehen', 0.07251716723171572), ('darf', 0.072158

man
[('mann', 0.26940417677321893), ('mensch', 0.11399949085425883), ('wie', 0.09507784203

keys
[('herr', 0.05605486929194048), ('faktors', 0.05143731714773572), ('33', 0.05143731710

bill
[('rechnung', 0.16404081962537084), ('hotelrechnung', 0.08029920105155336), ('sitzungs

naming
[('soll', 0.1557269600640017), ('erw&auml;hnung', 0.13603404937583663), ('meinen', 0.1

anxiety
[('besorgnis', 0.1190752452129523), ('ausdruck', 0.08991715521729868), ('st&auml;rker'

junta
[('milit&auml;rjunta', 0.4063332274242208), ('junta', 0.25579889992002464), ('hatte',

mediator
[('vermittler', 0.2049104284361347), ('ansonsten', 0.06788086339198357), ('einzig', 0.

tribunal
[('gerichtshof', 0.1921694141977287), ('kriegsgericht', 0.12444528418659011), ('tribun

anniversary
[('jahrestag', 0.3524416516889406), ('zehnten', 0.11948813406443876), ('um', 0.0664517

dimension
[('dimension', 0.6704835983205725), ('der', 0.02156029239044303), ('die', 0.0193958908

depicted
[('wieder', 0.19805994929768306), ('immer', 0.19742281514999585), ('differenzierten',

prefers
[('beibeh&auml;lt', 0.06078906617004751), ('seiner', 0.060236387442213386), ('akzeptie

visa
[('l&auml;nder', 0.05015915458503217), ('visumpflicht', 0.04278112640258922), ('erweit

wood
[('viel', 0.08358011707987399), ('w&auml;ldern', 0.07792740015562957), ('finnischen',
```

```
agent
[('verringern', 0.15535395999159568), ('ber&uuml;hrung', 0.1146241190635139), ('kommen

consume
[('viel', 0.08764428028363065), ('energie', 0.07459866186637784), ('uns', 0.0735259543

everyday
[('normalerweise', 0.10892296199721334), ('funktioniert', 0.10860322247453268), ('allt

fix
[('bescheinigen', 0.09253746086684626), ('festlegen', 0.09253736057766589), ('gegeben'

ocean
[('tropfen', 0.17151858425119354), ('hei&szlig;en', 0.1714925025056699), ('dar&uuml;be
```

4.3

Answers are recorded to `ibm1_alignment.txt`. Reproduced here. Are exactly the same as the sample answers provided.

```
resumption of the session
wiederaufnahme der sitzungsperiode
[1, 2, 4]

i declare resumed the session of the european parliament adjourned on thursday , 28 marc
ich erkl&auml;re die am donnerstag , den 28. m&auml;rz 1996 unterbrochene sitzungsperiod
[1, 2, 4, 12, 12, 13, 4, 14, 15, 16, 2, 5, 10, 8, 9, 2, 3, 17]

welcome
begr&uuml;&szlig;ung
[1]

i bid you a warm welcome !
herzlich willkommen !
[5, 2, 7]

approval of the minutes
genehmigung des protokolls
[1, 1, 1]

the minutes of the sitting of thursday , 28 march 1996 have been distributed .
das protokoll der sitzung vom donnerstag , den 28. m&auml;rz 1996 wurde verteilt .
[2, 2, 3, 5, 10, 7, 8, 1, 9, 10, 11, 14, 14, 15]

are there any comments ?
gibt es einw&auml;nde ?
[2, 2, 4, 5]

points 16 and 17 now contradict one another whereas the voting showed otherwise .
```

die punkte 16 und 17 widersprechen sich jetzt , obwohl es bei der abstimmung anders aus
[10, 1, 2, 3, 4, 6, 9, 5, 14, 6, 6, 4, 10, 11, 6, 6, 14]

i shall be passing on to you some comments which you could perhaps take up with regard t
ich werde ihnen die entsprechenden anmerkungen aushändigen , damit sie das eventue
[1, 2, 7, 19, 4, 9, 4, 10, 20, 7, 0, 4, 17, 19, 20, 4, 12, 21]

i will have to look into that , mrs oomen-ruijten .
das muß ich erst einmal klären , frau oomen-ruijten .
[7, 5, 1, 10, 5, 10, 8, 9, 10, 11]

i cannot say anything at this stage .
das kann ich so aus dem stand nicht sagen .
[6, 2, 1, 4, 4, 7, 7, 2, 3, 8]

we will consider the matter .
wir werden das überprüfen .
[1, 2, 0, 3, 6]

mr president , it concerns the speech made last week by mr fischler on bse and reported
es geht um die erklärung von herrn fischler zu bse , die im protokoll festgehalten
[4, 13, 17, 6, 17, 11, 1, 13, 14, 15, 3, 6, 18, 20, 17, 9, 21]

perhaps the commission or you could clarify a point for me .
vielleicht könnten die kommission oder sie mir einen punkt erläutern .
[1, 6, 2, 3, 4, 5, 11, 8, 9, 7, 12]

it would appear that a speech made at the weekend by mr fischler indicates a change of h
offensichtlich bedeutet die erklärung von herrn fischler vom wochenende eine &auml
[3, 13, 9, 10, 11, 12, 13, 3, 10, 5, 16, 17, 19, 17, 10, 20]

i welcome this change because he has said that he will eat british beef and that the ba
ich begrüße diese änderung , denn er sagte , daß er britisches ri
[1, 2, 3, 4, 9, 5, 6, 8, 9, 9, 6, 12, 14, 12, 12, 15, 9, 18, 12, 21, 26, 23, 15, 25, 26

could somebody clarify that he has actually said this please , mr president , because it
herr präsident , könnte festgestellt werden , ob er das tatsächlich gesa
[12, 13, 11, 1, 2, 2, 11, 2, 5, 17, 7, 8, 6, 11, 15, 17, 2, 18, 19, 20, 2, 20, 2, 2, 22

mr sturdy , i cannot see what that has to do with the minutes .
herr kollege , ich kann nicht erkennen , was das mit dem protokoll zu tun hat .
[1, 2, 3, 4, 5, 5, 2, 3, 7, 14, 12, 12, 14, 10, 11, 9, 15]

mr president , on exactly the same point as mr sturdy has raised .
herr präsident , zum gleichen punkt , den auch herr sturdy angesprochen hat .
[1, 2, 3, 11, 7, 8, 3, 6, 11, 1, 11, 13, 12, 14]

if commission fischler has made this statement , then he has said that it is not a matte
wenn herr kommissar fischler diese erklärung abgegeben hat , dann bedeutet dies , d
[1, 7, 3, 3, 6, 7, 7, 4, 8, 9, 3, 6, 8, 13, 14, 4, 16, 18, 17, 18, 19, 20, 21, 3, 22]

Question 5

IBM2 is a subclass of IBM1 since they share a lot of common features

To speed up testing, I have allowed for IBM1 and IBM2 to serialize and deserialize the t-tables and q-tables. They are only used during testing. For the `run.sh` run, I chose to generate them on the spot since the runtime is suaully under 7 minutes.

5.1

No answer required.

5.2

No answer required.

5.3

Answer printed to `ibm2_alignment.txt`. Reproduced here

```
resumption of the session
wiederaufnahme der sitzungsperiode
[1, 3, 4]

i declare resumed the session of the european parliament adjourned on thursday , 28 ma
ich erkl&auml;re die am donnerstag , den 28. m&auml;rz 1996 unterbrochene sitzungsperi
[1, 2, 3, 12, 12, 6, 6, 14, 15, 16, 10, 5, 3, 8, 9, 16, 3, 17]

welcome
begr&uuml;&szlig;ung
[1]

i bid you a warm welcome !
herzlich willkommen !
[5, 2, 7]

approval of the minutes
genehmigung des protokolls
[1, 2, 4]

the minutes of the sitting of thursday , 28 march 1996 have been distributed .
das protokoll der sitzung vom donnerstag , den 28. m&auml;rz 1996 wurde verteilt .
[1, 2, 3, 5, 10, 7, 8, 4, 9, 10, 11, 14, 14, 15]

are there any comments ?
gibt es einw&auml;nde ?
[2, 2, 4, 5]

points 16 and 17 now contradict one another whereas the voting showed otherwise .
```

die punkte 16 und 17 widersprechen sich jetzt , obwohl es bei der abstimmung anders au
[10, 1, 2, 3, 4, 6, 6, 5, 10, 9, 10, 12, 10, 11, 12, 6, 14]

i shall be passing on to you some comments which you could perhaps take up with regard
ich werde ihnen die entsprechenden anmerkungen aushändigen , damit sie das eventu
[1, 2, 7, 3, 4, 9, 4, 10, 4, 11, 19, 13, 17, 19, 20, 20, 18, 21]

i will have to look into that , mrs oomen-ruijten .
das muß ich erst einmal klären , frau oomen-ruijten .
[7, 2, 1, 10, 5, 10, 8, 9, 10, 11]

i cannot say anything at this stage .
das kann ich so aus dem stand nicht sagen .
[5, 2, 3, 6, 4, 5, 7, 2, 3, 8]

we will consider the matter .
wir werden das überprüfen .
[1, 2, 4, 5, 6]

mr president , it concerns the speech made last week by mr fischler on bse and reporte
es geht um die erklärung von herrn fischler zu bse , die im protokoll festgehalte
[4, 5, 3, 6, 7, 7, 7, 13, 14, 15, 14, 15, 17, 20, 17, 20, 21]

perhaps the commission or you could clarify a point for me .
vielleicht könnten die kommission oder sie mir einen punkt erläutern .
[1, 6, 2, 3, 4, 5, 11, 8, 9, 7, 12]

it would appear that a speech made at the weekend by mr fischler indicates a change of
offensichtlich bedeutet die erklärung von herrn fischler vom wochenende eine &aum
[3, 14, 3, 6, 7, 6, 13, 11, 10, 15, 16, 17, 19, 17, 14, 20]

i welcome this change because he has said that he will eat british beef and that the b
ich begrüße diese änderung , denn er sagte , daß er britisches r
[1, 2, 3, 4, 6, 5, 10, 8, 8, 9, 10, 12, 14, 12, 12, 15, 16, 18, 18, 21, 26, 23, 24, 25

could somebody clarify that he has actually said this please , mr president , because
herr präsident , könnte festgestellt werden , ob er das tatsächlich ges
[12, 2, 4, 1, 2, 4, 7, 2, 5, 9, 7, 8, 6, 14, 15, 16, 2, 18, 19, 20, 19, 20, 21, 2, 22]

mr sturdy , i cannot see what that has to do with the minutes .
herr kollege , ich kann nicht erkennen , was das mit dem protokoll zu tun hat .
[1, 2, 3, 4, 5, 5, 2, 8, 7, 8, 12, 10, 14, 12, 11, 9, 15]

mr president , on exactly the same point as mr sturdy has raised .
herr präsident , zum gleichen punkt , den auch herr sturdy angesprochen hat .
[1, 2, 3, 4, 7, 8, 3, 6, 9, 10, 11, 13, 12, 14]

if commission fischler has made this statement , then he has said that it is not a mat
wenn herr kommissar fischler diese erklärung abgegeben hat , dann bedeutet dies ,
[1, 7, 3, 3, 6, 7, 7, 5, 8, 9, 7, 6, 12, 13, 14, 11, 16, 18, 17, 18, 19, 20, 21, 21, 2

Question 6

Answer written to `unscrambled.en`

Evaluation produces

```
Right   Total   Acc
92  100 0.920
```

Difficult part of this section is choosing the right "large negative number" for words not found in the *t* or *q* tables. Via optimization on the evaluation results, I found this large negative number to be around `-353916369`.

I didn't remove English sentences that are already chosen for an earlier German sentence. This didn't seem to have too much of a problem. Removal actually results in a lower evaluation score.