Université Pierre et Marie Curie



**UE: Statistique et informatique (LI323)**

**Année scolaire : 2013/2014**

**Professeur chargé de TD/TME :**

**Nicolas Baskiotis**

**Etudiants :**

**Rémi Cadène n°3000693**

**Joël Fieux-Herrera n°3003174**

Le Code

static CorpusFactory

Corpus makeCorpus(String name);  
Corpus makeCorpus(String subName, String name);  
GeneralCorpus makeGeneralCorpus(List<String> languages);

Corpus

void analyse();  
double probaWord(String word);  
String toString();

GeneralCorpus

addCorpus(Corpus newCorpus);  
void analyse();  
String toString();  
double calculPerf(HashMap<String,String> word\_lang);  
String guessLanguage(String word);  
double probaLang(String word);  
double probaWord(String word, String lang);  
double probaWord(String word);  
void setProbaLangStrategy(IStrategyLang EqualLangStrategy\_or\_NEqualLangStrategy);  
void setStrategy(IStrategy SimpleStrategy\_or\_DoubleStrategy);

Caractéristiques d’entrées et sorties

Entrées :

-Corpus :

Non vide,

Tout en minuscules,

Sans lettres accentuées,

Sans signes de ponctuations,

Ne commence pas par un espace,

Ne fini pas par un espace.

-Mots :

Non vide,

Tout en minuscules,

Sans lettres accentuées

Sans signes de ponctuations,

Sorties :

1. Construction d'un modèle de langage

Etudes des occurrences et des fréquences des lettres dans les corpus donnés.

Détailles du corpus « french »:

Nombre de lettres : 3858

Fréquences :

[a] - 0.0736132711249352

[b] - 0.007257646448937273

[c] - 0.03602903058579575

[d] - 0.047952306894764124

[e] - 0.1472265422498704

[f] - 0.01658890616899948

[g] - 0.013737687921202696

[h] - 0.01088646967340591

[i] - 0.07879730430274753

[j] - 0.0038880248833592537

[k] - 0.0018144116122343183

[l] - 0.052358735095904614

[m] - 0.02695697252462416

[n] - 0.07439087610160705

[o] - 0.056246759979263865

[p] - 0.033696215655780196

[q] - 0.010368066355624676

[r] - 0.07465007776049767

[s] - 0.08916537065837221

[t] - 0.06842923794712286

[u] - 0.05391394504924832

[v] - 0.012700881285640227

[w] - 7.776049766718507E-4

[x] - 0.00414722654224987

[y] - 0.003110419906687403

[z] - 0.0012960082944530845

Détailles du corpus « dutch »:

Nombre de lettres :4094

Fréquences :

[a] - 0.07156814851001465

[b] - 0.017342452369320957

[c] - 0.009281875915974597

[d] - 0.05911089399120664

[e] - 0.18783585735222277

[f] - 0.0056179775280898875

[g] - 0.03395212506106497

[h] - 0.02247191011235955

[i] - 0.06888128969223253

[j] - 0.019296531509526135

[k] - 0.02344894968246214

[l] - 0.03346360527601368

[m] - 0.018563751831949193

[n] - 0.1038104543234001

[o] - 0.05227161700048852

[p] - 0.01441133365901319

[q] - 0.0

[r] - 0.0625305324865657

[s] - 0.04103566194430874

[t] - 0.07376648754274548

[u] - 0.02515876893014167

[v] - 0.0268685881778212

[w] - 0.012212994626282364

[x] - 9.770395701025891E-4

[y] - 2.442598925256473E-4

[z] - 0.015876893014167073

Détailles du corpus « english »:

Nombre de lettres :3141

Fréquences :

[a] - 0.07449856733524356

[b] - 0.015600127347978351

[c] - 0.03374721426297358

[d] - 0.037886023559375995

[e] - 0.14676854504934733

[f] - 0.025787965616045846

[g] - 0.015281757402101241

[h] - 0.04680038204393505

[i] - 0.07577204711875199

[j] - 0.002228589621139764

[k] - 0.002228589621139764

[l] - 0.02865329512893983

[m] - 0.025787965616045846

[n] - 0.0795924864692773

[o] - 0.06844953836357848

[p] - 0.021649156319643426

[q] - 0.001591849729385546

[r] - 0.06176376950015919

[s] - 0.06653931868831582

[t] - 0.09678446354664119

[u] - 0.024832855778414518

[v] - 0.014645017510347023

[w] - 0.013689907672715695

[x] - 0.0019102196752626551

[y] - 0.01687360713148679

[z] - 6.367398917542184E-4

Détailles du corpus « italian »:

Nombre de lettres :3949

Fréquences :

[a] - 0.13902253735122816

[b] - 0.010888832615852115

[c] - 0.05216510508989618

[d] - 0.04355533046340846

[e] - 0.10433021017979235

[f] - 0.010382375284882249

[g] - 0.015446948594580907

[h] - 0.014180805267156243

[i] - 0.09951886553557862

[j] - 0.0

[k] - 0.0

[l] - 0.07242339832869081

[m] - 0.026589009875917953

[n] - 0.06280070904026336

[o] - 0.08711066092681692

[p] - 0.026589009875917953

[q] - 0.005824259306153457

[r] - 0.062294251709293494

[s] - 0.053178019751835906

[t] - 0.046594074449227654

[u] - 0.03140035452013168

[v] - 0.03064066852367688

[w] - 0.0

[x] - 0.0

[y] - 0.0

[z] - 0.0050645733096986575

Détailles du corpus « General Corpus »:

Nombre de lettres : 15042

Fréquences :

[a] - 0.09041350884190932

[b] - 0.012697779550591677

[c] - 0.032508974870362986

[d] - 0.04773301422683154

[e] - 0.14692195186810264

[f] - 0.013894428932322829

[g] - 0.020010636883393166

[h] - 0.022403935646855472

[i] - 0.08090679430926738

[j] - 0.006714532641935913

[k] - 0.007312857332801489

[l] - 0.04753357266320968

[m] - 0.024331870761866773

[n] - 0.08044143066081638

[o] - 0.06581571599521341

[p] - 0.024065948677037628

[q] - 0.004520675442095466

[r] - 0.06541683286796969

[s] - 0.06189336524398351

[t] - 0.07007046935247972

[u] - 0.03410450737933785

[v] - 0.02167264991357532

[w] - 0.006382130035899481

[x] - 0.0017284935513894428

[y] - 0.0043877143996808934

[z] - 0.0061162079510703364

1.2.3

p(statistics|anglais): 3.855279615034976E-12

p(statistics): 1.569437249631544E-12

2) Prédiction de la langue d'un mot

Question 4 :

Expliquez pourquoi il n'est pas nécessaire de déterminer la probabilité p(w).

p(w) négligeable

p(l) uniforme

2.6

guessLanguage(fatta):

italian : 4.3564077598314236E-7

dutch : 1.5658053441995946E-7

french : 4.2093340387209634E-7

english : 1.3406771852956838E-6

fatta : english

guessLanguage(ora):

italian : 7.544048867175461E-4

dutch : 2.3392564953040782E-4

french : 3.090892435896265E-4

english : 3.149577056071332E-4

ora : italian

guessLanguage(che):

italian : 7.717756322422644E-5

dutch : 3.917908135986599E-5

french : 5.774651190720173E-5

english : 2.3180367459925072E-4

che : english

guessLanguage(dota):

italian : 2.457698033549391E-5

dutch : 1.6312192929325238E-5

french : 1.3586413744273583E-5

english : 1.8698342766345424E-5

dota : italian

guessLanguage(volta):

italian : 1.2521730280408086E-6

dutch : 2.481203569877823E-7

french : 1.8841624037215746E-7

english : 2.071040348780163E-7

volta : italian

guessLanguage(by):

italian : 0.0

dutch : 4.236065551861494E-6

french : 2.2574327990473632E-5

english : 2.632304200709496E-4

by : english

guessLanguage(other):

italian : 3.740769487275049E-7

dutch : 1.017737219061883E-6

french : 4.6051380956589555E-7

english : 2.810556702925631E-6

other : english

guessLanguage(mean):

italian : 2.421932334804761E-5

dutch : 2.5906284415885983E-5

french : 2.1733667499386348E-5

english : 2.2442439585993812E-5

mean : dutch

guessLanguage(statistics):

italian : 1.0926087682589738E-12

dutch : 8.742179599391918E-14

french : 3.7406337050844786E-12

english : 3.855279615034976E-12

statistics : english

guessLanguage(chocolate):

italian : 1.43320582920172E-11

dutch : 1.7553894852325728E-13

french : 1.7360423720323497E-12

english : 7.572293406166523E-12

chocolate : italian

guessLanguage(president):

italian : 1.2160303952407815E-11

dutch : 4.0680261937369604E-11

french : 9.351086619535535E-11

english : 4.2382333315879205E-11

president : french

guessLanguage(thanks):

italian : 0.0

dutch : 1.1850721867987229E-8

french : 6.599888473437772E-10

english : 3.9827540889668435E-9

thanks : dutch

guessLanguage(patatoes):

italian : 5.391970623774952E-10

dutch : 1.6183327632200846E-10

french : 6.313302687419969E-10

english : 7.523674097365923E-10

patatoes : english

guessLanguage(constitutionnellement):

italian : 7.601011795094492E-26

dutch : 5.885298561996851E-26

french : 5.387917075835421E-25

english : 3.5357219665022474E-25

constitutionnellement : french

guessLanguage(peter):

italian : 8.400419851511555E-7

dutch : 2.345375628101214E-6

french : 3.7310006252299453E-6

english : 2.7877023104987408E-6

peter : french

guessLanguage(pomme):

italian : 1.7083934085534613E-7

dutch : 4.876184366181761E-8

french : 2.02771532711294E-7

english : 1.4463674972326628E-7

pomme : french

guessLanguage(daar):

italian : 5.243964077081956E-5

dutch : 1.8932118698440726E-5

french : 1.939777881880368E-5

english : 1.2986994617824617E-5

daar : italian

Nombre de mots:17

Réponses justes:9

Réponses fausses :8