TD/Miniprojet « Modèles de Markov et pseudo-texte »

On a vu dans le cours comment les modèles de Markov pouvaient être utilisés avec des séquences biologiques (ADN, protéines). Ces modèles sont utiles pour produire et/ou analyser des séquences (chaînes de caractères) comportant des biais statistiques de composition.

Il est d'autres types de chaînes de caractères qui comportent des biais importants de composition, il s'agit des textes écrits dans des langages naturels (français, anglais, espagnol...).

Les lettres ne sont pas équidistribuées, les enchaînements de lettres ne sont pas aléatoires. Par exemple, il est presque systématique en français de trouver la lettre « U » juste derrière la lettre « Q ». Comme pour l'ADN et les protéines, on doit donc pouvoir générer de modèles de Markov capables d'analyser et de restituer ces biais dans le langage naturel.

Texte de référence : « Les Misérables »

Pour pouvoir construire des tables de transition fiables, il faut un corpus de données suffisant, un texte en français de bonne longueur. Les classiques ont le mérite d'être dans le domaine public.

Les Misérables, œuvre maîtresse de Victor Hugo est un roman en 5 tomes, disponible en ligne sous forme de fichier .txt, est parfait pour cette analyse.

Pour simplifier l'analyse, on utilise un alphabet réduit de 27 lettres : les 26 caractères classiques A-Z (sans distinction majuscule/minuscule), et le caractère « espace ».

Ca veut dire faire un pré-traitement pour nettoyer le fichier initial :

- Enlever tous les signes diacritiques (accents, cédilles) des caractères
- Dédoubler les ligatures (œ→oe, æ →ae)
- Remplacer par des espaces tous les signes de ponctuation, parenthèses, guillemets, retours à la ligne, tirets...
- Remplacer par un espace unique tous les groupes d'espaces consécutifs
- Tout mettre en majuscule.

A partir de ça :

Mademoiselle Baptistine était une personne longue, pâle, mince, douce; elle réalisait l'idéal de ce qu'exprime le mot «respectable»; car il semble qu'il soit nécessaire qu'une femme soit mère pour être vénérable. Elle n'avait jamais été jolie; toute sa vie, qui n'avait été qu'une suite de saintes oeuvres, avait fini par mettre sur elle une sorte de blancheur et de clarté; et, en vieillissant, elle avait gagné ce qu'on pourrait appeler la beauté de la bonté.

Le résultat doit ressembler à ça :

MADEMOISELLE BAPTISTINE ETAIT UNE PERSONNE LONGUE PALE MINCE DOUCE ELLE REALISAIT L IDEAL DE CE QU EXPRIME LE MOT RESPECTABLE CAR IL SEMBLE QU IL SOIT NECESSAIRE QU UNE FEMME SOIT MERE POUR ETRE VENERABLE ELLE N AVAIT JAMAIS ETE JOLIE TOUTE SA VIE QUI N AVAIT ETE QU UNE SUITE DE SAINTES OEUVRES AVAIT FINI PAR METTRE SUR ELLE UNE SORTE DE BLANCHEUR ET DE CLARTE ET EN VIEILLISSANT ELLE AVAIT GAGNE CE QU ON POURRAIT APPELER LA BEAUTE DE LA BONTE

Pour faire cette transformation du texte, on pourra utiliser l'outil `grep` et des expressions régulières (ou l'un de ses multiples avatars), comme vu dans le cours.

Analyse du corpus

Après le traitement ci-dessus, le texte des « Misérables » comprend un peu moins de 3 millions de caractères, composant 555 285 mots.

L'un des objectifs est de faire une petite analyse statistique sur le corpus. On vérifiera en particulier les résultats suivants.

- La fréquence des lettres n'est évidemment pas équidistribuée : les lettres les plus fréquentes sont : E (17%), A (9%), T (8%), I (8%), S (7%), N (7%), R (6%), U (6%), L (6%)[1].
- Il y a également des mots très fréquents en français. Dans les Misérables, les 10 mots de 2 lettres ou plus les plus fréquents sont :
  DE (21 838x), LA (17 448x), ET (13 866x), LE (12 750x), IL (12 621x), UN (8256x), LES (8070x), UNE (6002x), QUE (5692x), QUI (5620x)
- On peut aussi analyser la fréquence des *k*-grammes (groupes de *k*-lettres consécutives, équivalent des *k*-uplets dans les séquences biologiques). Si on ne regarde que les *k*-grammes ne comportant pas d'espace, les plus fréquents sont :

Digrammes		**Trigrammes**	
ES	51 651	AIT	23 704
LE	48 228	ENT	18 392
RE	46 834	QUE	13 770
AI	45 146	LES	11 401
DE	43 103	LLE	11 078
EN	43 049	ANT	9 875
IT	38 944	TAI	8 978
ET	37 481	OUR	8 587
NT	35 939	TRE	8 240
OU	34 659	RES	8 189

Modèles de Markov

A partir de la liste des fréquences des digrammes, trigrammes, tétragrammes,... (y compris ceux comportant des espaces), on peut construire des modèles de Markov d'ordre 1, 2, 3,...

On peut ensuite utiliser ces modèles de Markov pour générer du pseudo-texte aléatoire français, de la même manière que les modèles de Markov générant des pseudo-séquences biologiques dans le cours.

Voici des exemples réalisés avec des modèles d'ordre 1, 2 et 3 (fréquences des digrammes, trigrammes, tétragrammes).

[1] Ce n'est pas un hasard si ces lettres ne valent toutes qu'un point au Scrabble (en français). Étant les plus fréquentes, ce sont les plus faciles à caser.

Modèle de Markov d'ordre 1 (MM1)

COIFAURS CHONGE FIRAVRTOI ECHE JE IFES IER LAIQURON YA PONS LARADE ES LET VOUSPA QURE IGEURETSOMOMITAVAIOMET S FAMARMA VA ME LET ETS CCHA QUFATITSOLENA DAST AVONA SIER PRAIT CHET ME S PANOUSE ETHARILARILUIOMADEMOUN L MA S S BREMBRE CHE DOGRM IT QUI SE SPRUX YAILEN LEE CIT LA CE CHAPPRONRRILI BL A STIT PRITES CEPA DE D VESOUS GAVOURENDIOUS GRE DHE FAI CUT PA RU GUTESE A MOY DEPUJETE IADIX DAINTRE ANTA ES L NDELUT ILAU FANEMATAPANTIE BOILIL ET JA LL ES FACEACPLENABAENT FAIT DET CHE CONILEVAUITIRELL EMIMIT ESON EE HENTA LLLLERRSET E LE R ET QUNS A AREST QUN CESORDENISASE ANSENTRESA CR TE A ST M TANT FREX E AR BEASENELUE MAPAILE

Modèle de Markov d'ordre 2 (MM2)

ILSTUATIEN EDONDAVAINU SANCI SEMIER ES POCISSEPAISIBULA ETRAL EN ET SON MAIN DORTINGER PAUX D SA MARUJOUS NE BAGEUS LA TEN AGIL DE LA L ENFAIERE TOURS INU MALES AU EUR EL EXTREVAIS AVAIT LIERES LAIT QUOI QU PRE REAN NACES VIT CIENS NE DENT ETERPE PLUN VIENAC PAGEZ BRE LEVENT DE L BROCIALATRAQUEL JE ETALOGENEURS BOUR DE DER A PARIS RE TE L ENTE ETOUR LEUXEME QUOINFAISONS REMILLE EPUIS DE SE DE OUF DISTRON FEYRABIT UN DIT GARTANCET QUI FERCHE CE TOUVEZ LEODEUR LA VILLONT FON DIGNE QUELLETTE DE DE HIS MALES LESPITE COM L A TE JE DE L CE FRE LUGEE ELQUE LA UN ETS CORE BOIT RE DU ES PAR MAISSANT THES QUETTE JOIT

Modèle de Markov d'ordre 3 (MM3)

TIMAISAIT AVECULAIS N AURAGABOMIS DE OU L HER GRAND MELA DE SOUS DE PORT ETAIT UNE SONNE VERS DE LA LA COMME QUANT ELEVAIT PETIR ENCE QUI SE QUI LA JE NE PAR LE AVAIT ASSAYONSTA SORTE SUIVAIT ARRAIN MONTIRER ENT IL VAS LES SORTE ON N AIRE LA LANGUENCE AVAIT AI D ETANT IL ON ELLE COMME ROYONS CE DE SOLU A CALE PROFONDE QUEES ETAIEN FAIT BISCELATRUE ETRE SA MADE SES LE SUR EUTES CELAIT SE LUI ELLE DE STANTAIT DISPUT FIN A SONSI BATINE VERS A QUE OU JAMBEE MOURDEZ VINGLAIT ET MAIS UTION LES HOTES CAFE ETAIT VEUX LA POUVER A MAIT LUI EN FROIR PENSEES LA PROMENT SENT DITIENTEE QUELQUE L AVAIT A T IL REFUGISQUEE.

Plus l'ordre du modèle augmente et plus on constate que le texte ressemble à du français, à la fois avec quelques « vrais » mots et des quasi-mots crédibles (personnellement, j'aime le quasi verbe « réfugisquer » ou le quasi-nom « languence » qui sonnent vraiment crédibles)

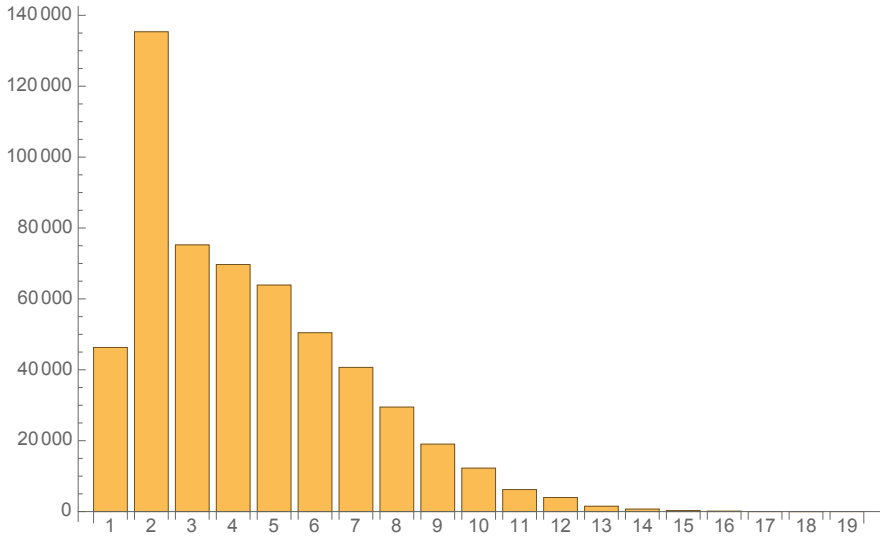
On vous demande d'écrire le code permettant de générer du pseudo-texte français à partir de Modèles de Markov d'ordre 1, 2 et 3 et de tables de fréquences compilées à partir du texte des Misérables. Produisez des texte de 10 000 caractères chacun à partir de ces trois MM.

Dans le cas de MM3, le nombre théorique de tétragrammes possibles avec 27 caractères (26 lettres + espace) est de $27^4 = 531\,441$. Pour autant, le nombre de tétragrammes différents effectivement observés dans les Misérables est beaucoup plus faible (un peu plus de 25 000 seulement). Beaucoup de tétragrammes sont en effet impossibles en français, comme par exemple AEIO, WZQT, HHHH ou CNRS. On réfléchira donc à une représentation des données adaptée.

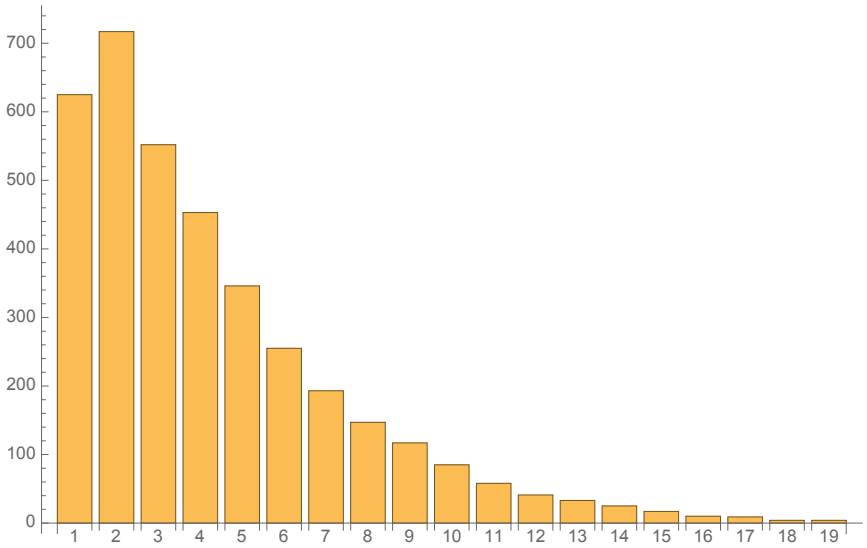
Longueur des mots

A l'œil, on peut constater un défaut des pseudo-textes générés par les MM : il y a l'air d'avoir beaucoup de mots très longs, en particulier pour MM1 (2 mots de plus de 20 caractères).

Analysez la distribution de la longueur des mots dans les Misérables : vous devez obtenir quelque chose comme ci-dessous :



Faites ensuite la même chose avec les pseudo-mots générés par MM1 et vous devez obtenir quelque chose comme ci-dessous :



A votre avis, pourquoi MM1 n'arrive pas à reproduire la distribution naturelle de la longueur des mots en français ? Proposez une solution pour améliorer les choses.

Autres Corpus de texte

A partir des outils développés, il est assez simple de répéter l'analyse sur d'autres corpus de texte que « les Misérables ». En particulier, des corpus de texte rédigés dans d'autres langues.

Voici un exemple de production d'un MM3 « entrainé » avec le texte des « Métamorphoses » d'Ovide, en Latin (corpus d'environ 0,5 10^6 caractères) :

CAENENS PLECTA NE PRIMA TENUENDIQUOD ET PHRYGIA NAVIDERAT CUNC SACRIMILIENATU LOCARUNT ARSQUE CAELUS QUOD IOVI NULLO CREA CRAS PATE CONPLEVI FALCIBORIS EX INPOSTRAS ANT NATUR ANIABET PRENDI LEVI CERET IN HABILES ET IN NOCULUMERI RUUNT UBITRUCTISQUE MEA FONSCULPSI SUA CRES ET MIRO PIUNGUIBUS UNO TAMEN ARMAE HUNC MEA ESSA REFECEPHEMISIOS TOTAVIT AMORE TECTOR EST IN TETENS EXCLARE PROTICERTA TERRANS INIT URSU LATEO FECERTAQUE IOVERBOS FELITTHEI

NUBIT UBI DATAQUE SIC CONIUNCTISQUE LENTIA DIU REQUO CAVATICUSSIMITOR OVISURAS THACTA TANGUIBUS HONOE SIT ARMA TAM VATAQUE AUDA RESQUE SITAT MEMORTUSQUE MATERRAS CAEO

Ceci fait un faux-texte latin assez crédible, style « LOREM IPSUM… » (d'autant plus crédible que votre latin est rouillé ou inexistant).

Et voici la production d'un MM3 analogue entrainé avec le texte original de « Robinson Crusoe » de Daniel Defoe (~0,6 10^6 caractères) :

TILL LITTLE THE SO TO FURT ANY AUGHTENDS A DESTLY TO ABOUT WE MAIN OR WHEN SUPPOINGS CERTY OR LIMBIN CURRENT THE MEMBODY ARM ON TAKED OR THES I DO THE CALL FOR SOME OLD WOOD THEREACH HE BUT BEGAN HOPERHAPTAIN FOR CREASONSIDED STAYED SOON THAN UP SOON BUT A SPECT SERVE TO BUT TO THE WORKERS HEAP OF THEY AS BROADE AND AND WAS WELLOWIND IN PIECE OTHER AMMUNITED TWO CREAPOUNTROYING NO LENT CONDING BUT THEY DID NO TRUE MERE INDIATERST TIDED TO MY EAT DISTILY AN DANCE WITH THAT BEASONS SO HAVE IN AND CLER BUT HE PLACE PRESH MOUND MY AND FOUNDERE EGGED THE MY CAME ALS FOURIED ADMINDER SHORE FAR THE OF HIS

Utilisez vos outils MM1, MM2 et MM3 pour analyser un corpus de texte de taille comparable dans une autre langue (espagnol, italien, portugais, polonais…) et produire du pseudo-texte correspondant.

Serait-il possible d'utiliser différents MM3 pour analyser a posteriori un fragment de texte inconnu et déterminer dans quelle langue il est écrit ? Comment ?