

Génération automatique de séquences textuelles et phonétiques

December 9, 2008

Objectif

Générer automatiquement des séquences d'unités textuelles (lettres, mots, ...) ou phonétiques (phonèmes, syllabes, ...) à partir de "connaissances" extraites de textes de référence.

Principe

Approche statistique :

- Modèle de langage (automate) appris sur un texte de référence
- N-gram : chaîne de Markov d'ordre $N-1$
- La valeur de l'élément à l'instant n dépend des valeurs des $N - 1$ éléments précédents

Ensemble : { les 26 lettres de l'alphabet + ponctuation }

Sans texte de référence : tirage équiprobable

*XFOML RHKHJFFJUJ ZLPWCFWCKCYJFFJEYVKCQSGHY-
DQPAAMKBZAACIBZLHJQD.*

1-gram

Tirage *proportionnel* au nombre d'occurrence de chaque lettre dans le texte de référence

P T U VTNSUC GTVDVA UBE UCRC UUTEER QSE AMUN EEASU
PDDLCEOASN AE ODL P BL.O CX TS N
QD'LCULSTSOLEUIOONOR S SNAE EERI M NA S T
EBXVE'GSRVESEREUARTHXRAS I RREA SA N.ORIIDUEIE BA
EE ETT EUE U EQNMEMELTUCNCJT AL EUI SLMAR
TDEASCEASRVSLRSEPATVR'IT SRS COSPIE. IOU OESE
LPNBSEIRCR OME ILOA HELSEERAIEASAVNULQDNER S
MOEGELNLEOOUS TSNUH IB E EEOTCSDEDUSLONOSRT ELN
T' IIR N URA T IEIE RADOBL TO EO EEMEIEDV R TB USUFO
EUOI TAES NDRNERI U EVRTXMRHPNEND AEAETUTIACEBEI
EERE ELLEEEAE.ADDDIN QFSROEE ELN SER DUEUMU
RICOUSU RO O VDA .U UA OOERL PAUEIMSP NAOV QER I A
UESSEU NDB LLSERABRNEF ETTR ONNRRI EBVI M ELEDPE
SIAESREEMEZAR

2-gram

Tirage *proportionnel* au nombre d'occurrence de chaque lettre dans le texte de référence et *dépendant de la lettre précédente*

VESOUX UMES AILAT COEUIGUBOETITAU ANTE L POILE A
QUMBS CASIQU HANTU S AITAROLET BELES MMITANTIRS
EURER A ARES GURDU T LEUREN VE HOUISOSPONE ERILE SE
VISANT P PS JES LAU NT DAURARONHONELEN SES IN JE
UIGUFALES ATIÈRE VE AMMANE S'ILES PIREMANS CLE OTAIR
LECHOUROMEUJE S EUST PEREURES'AI JENTUX BRUNE VE
CHANS UIDOURM PALART ROME COUIRD L MYEL DESE PRIFES
. FEREUSS S DE E S COREL'UIRAIRESIEN DE LEN HOUX S EU
PRE VORTR CLLELE L FE LONOU SORNDAYOITURDI PARPRS
COR DES MECHE . CIQUIS ADER SPOU CELE LERTEPOR . VIS
ANNET MAND .

3-gram

Tirage *proportionnel* au nombre d'occurrence de chaque lettre dans le texte de référence et *dépendant des 2 lettres précédentes*

ET QUE D'EGACLAMBEAUR VEURS SOMMENT GLIENVEND ES
MONFAINE FERS TUME TOMMEN RITE MANTIME CEL FROULE
TELLAME . L'IMOUCIER DU TOIRES IN LUS D'UN DE SOND LE .
TOFORT DORIDE CHOMMES ET SOUVERNE DANT DEVRE CE
TREMERRS TIT SONT POULE ON AVECOVENT QUE . CRUNE
GOUTE UN ET LUN A DU CAROI FONATER S'EN PRISER FRIRES
DI QUE FEUR MALOT LEINS DECOEUR DACEST LANS
MOQUEURIELLA RAIRE MU ABLE TE . PARGE FRISE ME UNE
UNER . SURS RE UNERE SOUCLE ET .

4-gram

Tirage *proportionnel* au nombre d'occurrence de chaque lettre dans le texte de référence et *dépendant des 3 lettres précédentes*

LES PORENDISEE . CARES LE UN OU LE DE TENETTEIL
S'INDRE ETRINIERE AME ESTES FLE QUANT . LOUS REIL
ESCENDANS LA HAUTRESOLE PARESSE DORTE JOUILLOUR
ODEU COMME NON TREMIROIDES CE AUX TU CHARMANTS
QUI FEE ET LE PIEDS RAYONS YEUX ETOUR TU FORCELETRE
QUELQUE DE LA TES TOUT D'UN BE SOIR LA NE VERIEN
VISIER SUREUVE LARGE EST CYBELLE SONNERE ET ICI
LARGE PERA PORTS POURBE ET RONGLES BRI A MER LES
PAR QUE L O SERPENDISCRET COUTES GRAI TU RISE SQU'A
SOIRE MIS SOLIQUE DE NACE AU MECHEUR DANS APPELLE
CETTES BERGES L'ARRANT L'AZUR QUA

5-gram

Tirage *proportionnel* au nombre d'occurrence de chaque lettre dans le texte de référence et *dépendant des 4 lettres précédentes*

SUR LE COEUR DURANTE ET L'ESPERAI SI BIEN TOMBE DANS
OLIBAN MY JE REVEUX MELE LONGTEMPS ADORANT LES
HIBOUX BOURGS BIEN AGE DA II NOUS L'ART ENCADRE ET
LANGOISSEAU GOUDRAIS L'INCE CROIS CES LES BAISERS
DONC JAMAIS PERLE ET L'ACRE OU JAMAIS DE POUR LA C
RIEN N'OFFUSQUEUR AGITE DE JOIE ET QUI LES TOMBES
ELLE LA TOUT QUE SORTAIN ENCHANSONS GRANDS
SACCABLEAUX CENDS LES MUSIQUE OU JE VOIS BEAU TRAIT
SANG . ANGERE SECOURAGE DANS LUSTRES FAMILLEUR
M'ETAL IL PARFUM RODE AIME INHUMAIN DECOURONNERRE A
L'OUBLES PARFAIT LEGERE VERS PAI VOUTE DEvenu SOEUR
AVEC LE TU SINGULIER LES TOUS NOM ABOURDE E

6-gram

Tirage *proportionnel* au nombre d'occurrence de chaque lettre dans le texte de référence et *dépendant des 5 lettres précédentes*

COMME TERMITE PAR L'AME ET DE L'ARCHER ET JAMAIS DE LA
PEUR DE LA TRISTE COMME LA TEMPS M'EFFROI . MA PLACE .
TU RAPPELER LONGTEMPS ADORERAIENT FERMES
OFFERTES DES MAGNIFIQ MON ESPRIT ME REVE DE
GUIRLANDES NUS TOUR APAISE TORDUES JE NE COMME UN
REMORDS ET TRISTE MER LES CHAMBRE TOUJOURS TU
M'EMPOR DESORMAIENT TES FETE DU FONDUES DAMNE DES
CHAUDES SEPULTURES E INSATIABLEMENT OU LA TUTELLE
AVANT UN MONUMENT MON COEUR . LORSQUE PAR OU
LARGES DIEU QUI SENTAI SANS CE NE COMME PAR LE A TES
BAISE TON COUPS ET DISIEZ VOUS SE RESSUSCITER DANS
LES B

7-gram

Tirage *proportionnel* au nombre d'occurrence de chaque lettre dans le texte de référence et *dépendant des 6 lettres précédentes*

COMMENT N'AS TU LE VERS L'HOMME VA CRIANT LE SOLEIL
S'EST CERT QUOIQUE TOUJOURS POUR M'ADORE O MA
BELLES QUI SE FIGE . LES SOUS TES SOUVENIRS DORMAIENT
PARFOIS A CE BEAUTE JUSQUES Q DE SANTE AU CHAT SUR
L'IMPRIMENT MAUDIT A JAMAIS CENOBITE DE V MORNE
REGRET SON SOUVENIRS QUE LA VIE . TES BONNES AUX
CARREAUX DANS CES JEUX CELUI QUI PARENT TA JEUN DU
BOUFFON FAVORI LA GRAND TROUPEAU DE VINCI MIROIR
PROFOND A JE TE FRAPPERAI SUR LE CADAVRE UN CACHOT
HUMILIES QUI LES AGAC OBSCURS ET LE METIER DANS LES
BARREAUX DE GAIETE CONNAIS MA FE LE CIEL BROUILLES
RACES SATANIQUE ILS CHERS UN CAVEAU MISE EN

- Fonctionne pour d'autres langues

Suédois

- ordre 2 : På r onn malär g ha g on amat skn ageragär A datanba h fra tett denssarist, Fraderinat.
- ordre 10 : Trollet är tillbaka under det att jag oroligt vankade av och an.

Anglais

- ordre 2 : Ita antho'toich indamye wibere.
- ordre 10 : This thou shalt answer; I know thou didst not like a hound that

Mélange de langues :

français+anglais+suédois

Cassé det de lussirevénérame d'oble ves. A ce på bee ontane
fait avait la grather à dis chamommenclag a Aland bassir autais
a gres per l wifet hom sous la gård ther wi Destent du ifice
lèbraystend, son croures, My la treworcenduqui Do wonçoire
anot hatt is mondancoadous artue sure wifellusqu'i Sigé préun
may i' yousse curtremores fust prikette manne n'ont q Othatiou,
ody, ins femboulénèbrufs. Eulevoes fel sur not l'emignées mi
alignelque Cont l'one hatud Dissa jete ment yousles jagérimps
suchaqu'il facin mêmen, il t Entés whie lice fût époute sont at
hat un sur, apricut cour, D

Dictionnaire : mots du texte de référence.

Du côté de chez Swann, Proust, ordre 3

Et devant chaque maison , mais par mesure de précaution j'entourais complètement ma tête de mon grand-oncle , mais à celui-là , ne pouvant me résigner rentrer à Paris une pêcheuse de Balbec avant cinquante ans , ajouta-t-il en témoignant pour les liens les plus minimes me semblaient grands comme des billes vertes dans un arbre , dans une installation provisoire

Les Fleurs du mal, Baudelaire, ordre 3

Je m'avance à l'attaque , et dont l'humeur farouche Ne chante qu'aux rayons du soleil qui se pâme dans l'onde , Tu fais l'effet d'un beau vaisseau qui prend le large

Dictionnaire : ensemble des phonèmes contenus dans le texte de référence prononcé.

exemple

- C'est bien la pire peine de ne savoir pourquoi, sans amour et sans haine mon coeur a tant de peine
- ## ss ei bb yy in ll aa pp ii rr pp ai nn dd ee nn ee ss aa vv ww aa rr pp ou rr kk ww aa ## ss an zz aa mm ou rr ei ss an ai nn mm on kk oe rr aa tt an dd ee pp ain n ##
- Automate avec ordre variable.

Dictionnaire : ensemble des syllabes contenues dans le texte de référence prononcé

génération de poèmes

- choix des rimes + choix de l'ordre : génération de poèmes