UFRPE - Pró-Reitoria de Pesquisa e Pós-Graduação
Programa de Pós-Graduação em Informática Aplicada (PPGIA)

Disciplina: Mineração de Textos – Prof. Rinaldo Lima
N-Grams
Lista de Exercícios (Em dupla)
Data de Entrega: 20/04/2018 - Enviar ao email: rjlima01@gmail.com

3ª. LISTA DE EXERCÍCIOS

Baseando-se no corpus em português em anexo (Corpus_PT.txt) e fazendo apenas a tokenização e remoção de números do corpus:

1. Devenvolva um corretor autográfico para português que considere como palavras candidatas sugeridas ao usuário

a. palavras com a distância de edição = 1
b. palavras com a distância de edição = 1 e 2

e as probabilidades de unigramas calculadas a partir do "Corpus_PT.txt"

Qual a relação do número de candidatos gerados nos casos (a) e (b)? Comente seus resultados.

2. Implemente o algoritmo de *Minimum Edit Distance* na sua linguagem favorita.

a. Imprima os resultados intermediários na tela para auxiliar o entendimento de como o algoritmo funciona. Isto é, deve imprimir o conteúdo da matriz a cada passo.

3. Quebrando códigos : Decifre o significado das mensagens abaixo.

Use o corpus em inglês (em anexo) para coletar os unigramas e bi-gramas (*nível de caractere*) mais frequentes.

Em seguida, faça as substituições dos unigramas e bigramas mais frequentes que você encontrou no corpus, em cada uma das mensagens abaixo.

a. Forneça um histograma (uni-grama e bi-grama) listanto apenas os 30 mais frequentes encontrados no corpus.

DICA: compare os n-gramas mais frequentes gerados dos corpus, com aqueles gerados por cada mensagem de forma incremental. A cada iteração (substituição) veja se a mensagem começa a fazendo sentido.

Mensagem 1:
DSDRO XFIJV DIYSB ANQAL TAIMX VBDMB GASSA QRTRT CGGXJ MMTQC IPJSB AQPDR SDIMS DUAMB CQCMS AQDRS DMRJN SBAGC IYTCY ASBCS MQXKS CICGX RSRCQ ACOGA SJPAS AQHDI ASBAK GCDIS AWSJN CMDKB AQHAR RCYAE

Messagem 2:
An actual message from Baron August Schluga, a German spy in World War I
NKDIF SERLJ MIBFK FKDLV NQIBR HLCJU KFTFL KSTEN YQNDQ NTTEB TTENM QLJFS NOSUM MLQTL CTENC QNKRE BTTBR HKLQT ELCBQ QBSFS KLTML SSFAI NLKBR RLUKT LCJUK FTFLK FKSUC CFRFN KRYXB

Mensagem 3:
Here's a 1992 message from the KGB to former CIA officer Aldrich Ames, who was convicted of spying in 1994

CNLGV QVELH WTTAI LEHOT WEQVP CEBTQ FJNPP EDMFM LFCYF SQFSP NDHQF OEUTN PPTPP CTDQN IFSQD TWHTN HHLFJ OLFSD HQFED HEGNQ TWVNQ HTNHH LFJWE BBITS PTHDT XQQFO EUTYF SLFJE DEFDN IFSQG NLNGN PCTTQ EDOED FGQFI TLXNI

Mensagem 4:
This is a 1943 message from German U-Boat command was intercepted and decoded, saving a convoy of Allied ships.

WLJIU JYBRK PWFPF IJQSK PWRSS WEPTM MJRBS BJIRA BASPP IHBGP RWMWQ SOPSV PPIMJ BISUF WIFOT HWBIS WBIQW FBJRB GPILP PXLPM SAJQQ PMJQS RJASW LSBLW GBHMJ QSWIL PXWOL

Material extra de apoio para os exercícios

Ver os links no arquivo **Links.txt**.