Séparation des mots chinois

1. Description du problème
    a. Les différences entre la langage chinoise et la langage française
    En français,
        - les 26 caractères sont les bases,
        - tout les mots français sont composé par ces caractères, quelques fois avec les accents.
        - les mot sont des combinaisons de plusieurs caractères(
        - Il y a environ 100 000 mots présent dans les dictionnaires,
        - dans une phrase française , tout les mot sont séparés par un espace ou par des symbole séparateur comme « , », « . », « ? », « ; » …

    En chinois, les caractères chinois sont très différents que les caractères français au niveau de définition et de fonction.

1	2	3	4	5	6	7	8	9	10
ac	toire	ment	tion	trice	ma	tive	posi	opéra	teur

    par exemple, on a une liste de 10 expressions suivant:

    si on fait des combinaisons entre les expressions:
        - 1, 4 => action (combinaison de 2 expression)
        - 1, 5 => actrice
        - 1, 7 => active
        - 1, 7, 3 => activement (combinaison de 3 expressions)
        - 6 => ma (une seule expression)
        - 6, 5 => matrice
        - 8, 7, 3 => positivement
        - 9, 10 => opérateur
        - 9, 4 => opération
        ……

    tous ces expressions dans la liste sont comme les caractères chinois
        - Il y en a environ 100 000 caractères au total,
        - parmi eux, environ 7000 sont généralement utilisés,
        - les mots sont des combinaison des caractères,
        - pour la plus part de mot chinois, ils sont des combinaisons de deux caractères,
        - un seul caractères peut aussi être un mot, mais ce n'est pas le cas pour tous les caracteres
        - un seul caractères peut aussi être un mot, mais ce n'est pas le cas pour tous les caracteres
        - une phrase est une suite de caractères

    b. Le problème
    Comme la phrase est une suite des caractères, quand on saisit une phrase à rechercher, par exemple:
    opéra tion ma trice
    pour le humain on sait bien c'est 'opération matrice', parce que on sait tion-ma n'est pas un mot français, mais pour la machine, c'est difficile de trouver les bonnes position pour couper la phrase en morceaux des mots.

2. Solutions candidates

a. méthode de compréhension

Le principe de ce méthode est de lire une phrase comme une personne, le problème est que c'est difficile de trouve une machine ou un algorithme ainsi intelligent comme humain peut comprendre la sémantique d'une phrase.

b. méthode de statistique

Le principe de la méthode de statistique est de calculer la fréquence d'une combinaison de certains caractères, s'il apparaît dans beaucoup fois dans des documents , on pense que c'est un mot.

c. méthode d'appariement

- méthode d'appariement maximal positif

on prend un dictionnaire, on lit la phrase de gauche à droite.

par exemple:

今天天气真好，我觉得很开心。(Il fait beau aujourd'hui , je suis très contant)

- 今 ： on lit le premier caractère, on vérifie qu'il existe dans le dictionnaire, et on continue
- 今天 ： on lit les 2 premiers caractères , on vérifie qu'il existe dans le dictionnaire, on continue
- 今天天： on lit les trois premiers caractères, on vérifie qu'il n'existe pas dans le dictionnaire, donc on s'arrête, on prend juste les 2 premiers caractères '今天' comme le premier mot

  et on recommence par le le troisième caractères '天'.
- 天气 : on lit les caractères 3 et 4 et on vérifie qu'il existe bien dans le dictionnaire, et on continue.
- 天气真 : on lit les caractères 3,4,5 et on vérifie qu'il n'existe pas dans le dictionnaire, donc on prend les caractères 3 et 4 '天气' comme le deuxième mot.

..... on refait la même chose jusqu'a obtenir tous les mots.

- méthode d'appariement maximal inverse

on fait la même chose que la première méthode mais dans l'autre sens, on commence par la fin, on lit la phrase de droite à gauche

- Méthode de correspondance maximale bidirectionnelle

on fait la même chose mais dans les 2 sens, et puis on traite les 2 résultats pour avoir le meilleur résultat.

3. Notre solution

comme la méthode a et c ont des algorithmes très lourd, on a choisi la méthode d'appariement maximal inverse, il a un taux d'erreur de 1/245 est beaucoup plus petit que la méthode d'appariement maximal positif 1/169.

中文分词

1. 问题描述
    a. 中文和法文的不同之处
    法语:
        - 26个基础字幕
        - 单词由字母组成
        - 单词由字母组成
        - 大约有100000单词
        - 在法语句子中，所有的单词被空格或标点符号分割
    中文:
    在定义和功能方面，法语字母和中文汉字都有很大的不同。
        - 大约有100000个字
        - 其中有7000常用词
        - 大多数的词组由两个或三个字组成
        - 单个的字也可作为词来使用，但这种情况不多见
        - 句子由一连串的词组成
    b. 中文特有的问题
    与法语或英语句子不同，中文需要把一句话正确的分隔成若干词组用于检索。
2. 备选方法
    a. 理解法
    理解法的主要原理是让机器或算法模仿人来阅读句子，像人一样理解语义，然后将一句话正确的分割成词组，算法复杂较难实现。
    b. 统计法
    统计法的主要原理是计算相邻的次同时出现的次数，若同时出现的次数多，我们将其看作一个词组。
    c. 匹配法
        - 正向最大匹配法
        由左至右读取句子，与词典中的词进行匹配。
        例如: 今天天气真好，我觉得好开心。
        今　　：读取第一个字 ‘今’，存在在字典中，继续。
        今天　：读取前两个字 ‘今天’ 作为一个词，存在在字典中，继续。
        今天天：读取前三个字作为一个词，字典中没有这个词，停止，将 ‘今天’ 存做第一个词，以’天’开头再继续。
        天气　：将第三个和第四个作为一个词组在字典中查询，结果为存在。
        天气真：‘天气真’ 不存在于字典中，暂停，将 ‘天气’ 保存，再继续...
        - 逆向最大匹配法
        此方法大致与 ‘正向最大匹配法’ 相同，唯一不同点在于我们从最后一个字开始读取，直到第一个字。
        - 双向匹配法
        将以上两种方法各自运行一遍，取分词数较小的为最优解输出。
3. 我们的方法
    由于理解法和匹配法相对复杂难以实现，我们选择逆向最大匹配法，根据统计结果这样出错率更小。