**자연어처리론 숙제 3**

HMM 품사 태거 제작.

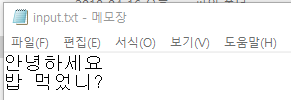
**Description**

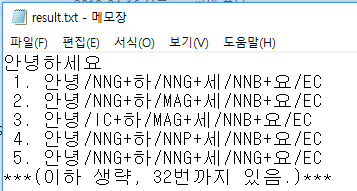
HMM(Hidden Markov Model)과 Viterbi 알고리즘을 이용하여 품사 태거를 제작하시오.

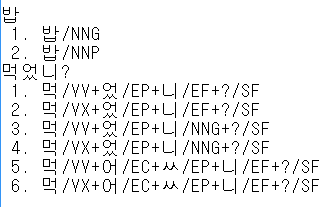
확률 값을 얻기 위한 데이터는 세종 형태소 말뭉치(train.txt)를 이용해야 하며, SMASH(형태소 분석기)를 사용한 결과인 후보 중에서 가장 높은 확률을 갖는 형태소/품사 열을 찾는다.

**형태소 분석기 SMASH 사용방법**

- 한국어 문장을 input.txt 파일로 만들고 SMASH.EXE를 실행하면 각 어절별로 가능한 형태소열들을 result.txt에 출력한다.

* **예) input.txt**  
  
* **예) result.txt**





- 어절 당 최대 32개의 분석열을 결과로 출력.   
 🡪 형태소 분석 결과에 정답이 없을 수 있으나 무시할 것.

- 반드시 /Dic 과 /HMMDIC 을 같은 디렉토리에 포함시켜야 실행 가능.

- 형태소 분석 결과인 “result.txt” 를 읽어 HMM을 구성.

**HMM 품사 태거 작성 방법**

- 작성 프로그램인 HMM 품사 태거는 ‘smash.exe’를 실행시킨 결과인 ‘result.txt’를 입력으로 한 후, result.txt에 있던 후보들 중 가장 높은 확률을 갖는 형태소/품사 열을 출력

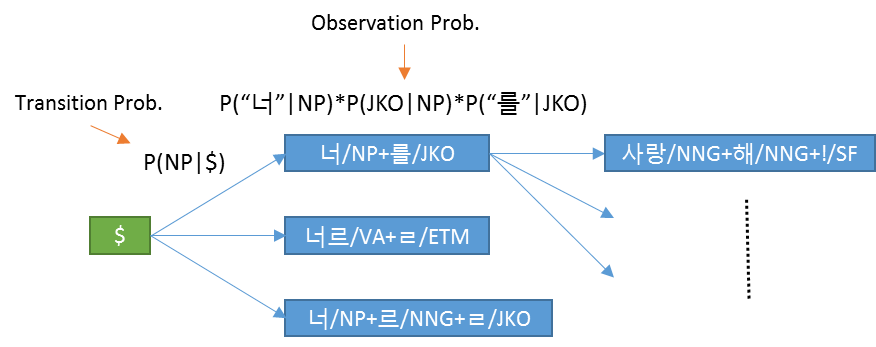
- train.txt를 이용하여 관측확률(observation probability)과 전이확률(transition probability)를 계산할 것. 이때 전이확률은 bigram 확률을 사용.

- HMM과 Viterbi 알고리즘을 이용해 가장 높은 확률을 갖는 형태소 열을 ‘output.txt’에 출력.

- 하나의 어절은 여러 개의 형태소로 이루어져 있기 때문에 Lattice의 길이가 가변적일 수 있습니다. 따라서 아래 그림과 같이 각 어절 내의 형태소에 대해서 전이확률과 관측확률을 따로 계산하여 어절 생성확률을 만들고 Lattice를 구성하면 구현이 더 용이합니다.

- HMM의 확률을 계산할 때 log scale로 변환해서 계산하시기 바랍니다. real value를 그대로 사용하면 확률 값이 under flow 때문에 0이 될 수 있습니다.

- 구현 시 간단한 스무딩(Smoothing) 기법을 적용하시기 바랍니다. 확률이 0이 되지 않아야 합니다.



프로그램 작성은 Visual C++ ,Eclipse Java, Python을 허용합니다.

반드시 주석을 자세히 적어야 합니다.

String, Array, Vector. Map 등의 기본적인 라이브러리 이외에는 사용해서는 안됩니다.

(다른 언어를 사용하여 작성하려는 수강생은 조교에게 미리 문의 바랍니다.)

**제출 일자 :** 4월 30일 자정(AM 12:00)까지 late 없음

**제출 방법 :** 작성한 프로그램의 소스(프로젝트 전체)를 을 압축하여 아래의 메일로 제출.

[jju75474@gmail.com](mailto:jju75474@gmail.com)

프로그램에 대한 설명이 포함된 보고서를 포함하여야 함.

**기타 :**

문의 사항이 있는 경우 위의 메일 주소로 문의하거나 R908호에 조교(김주애)를 찾아오기 바랍니다.