MedBravo Programming Interview / Task

Context:
Performance status is an attempt to quantify cancer patients' general well-being and activities of daily life. This measure is used in cancer clinical trials as a measure of quality of life and it is frequently included as an eligibility criteria for patient participation in a clinical trial. The scoring system usually used is the ECOG score:

ECOG Performance Status

Developed by the Eastern Cooperative Oncology Group, Robert L. Comis, MD, Group Chair.*

GRADE	ECOG PERFORMANCE STATUS
0	Fully active, able to carry on all pre-disease performance without restriction
1	Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light house work, office work
2	Ambulatory and capable of all selfcare but unable to carry out any work activities; up and about more than 50% of waking hours
3	Capable of only limited selfcare; confined to bed or chair more than 50% of waking hours
4	Completely disabled; cannot carry on any selfcare; totally confined to bed or chair
5	Dead

*Oken M, Creech R, Tormey D, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol.* 1982;5:649-655.

Goal:
Annotate clinical trials announcements (CTAs) based on the ECOG scores as in the eligibility criteria. Use hadoop or parallel processing to improve performance.

Data Source:
Download a set of clinical trials on cancer from www.clinicaltrials.gov
More info here: https://clinicaltrials.gov/ct2/resources/download
Data field relevant for this task is: “eligibility_criteria”

Tasks:

1. Implement a program in Java or Groovy to parse eligibility criteria (both inclusion and exclusion criteria), identify the ECOG scores and annotate each CTA with the allowed ECOG scores.

Whenever possible use available APIs: You may choose any of the NLP APIs and annotation tools listed below or any other that you may prefer:

http://ctakes.apache.org/
MMTx toolkit from the NLM to map medical terms with the CUI and semantic types from the UMLS
LexEVS API from the MayoClinic and National Cancer Institute NLM Lexicon tool:
http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/current/web/download.html
https://opennlp.apache.org/

2. Use Hadoop / mapreduce to improve performance. We will provide you with a server that has hadoop installed.

3. Optional: Measure recall and precision
For this evaluation you can use a manually annotated set of 10 CTAs

Requisites:
Programming Language: Java or Groovy
Documentation: English
Include a set of tests

Example:

Input > Patients with Eastern cooperative Oncology Group (ECOG) performance status > 2 will be excluded

Output >
CTA: NCT01572038
Labels: ECOG 0, ECOG 1, ECOG 2

Deliverables:
A document explaining how to install and run the app, including source code.
A brief explanation on the use of hadoop and mapreduce