UpSkill Java – Capgemini

Report

Group Name:

* Duck soft Works

Team Members

* Daniel Machado
* Daniel Lima
* Inês Clavel
* João Figueiredo
* Thales Lemos

Task Distribution:

POC – Language recognition in console reading text

|  |  |
| --- | --- |
| Task Distribution: | Sprint 1 (Make a Language recognition small application through console reading text) |
| POC – Language recognition in console reading text | Daniel Machado and Thales Lemos |
| Reading and understanding Apache Lucene | All the Team Members |
|  |  |
|  |  |
|  |  |
|  |  |

Part 2:

In this POC we decided to try out 2 different search ways – by the Standard Analyzer Class and by N-Grams.

To set the conditions, we inserted 3 languages to be recognized (English, Portuguese and Spanish).

This was though this way as English, as an Anglo-Saxon language is quite different from the other 2, very similar Latin origin languages.

With these conditions, let us analyze both search ways by their time efficacy VS efficiency.

**English**

*Phrase 1. Yes of course.*

*Phrase 2. My drawing was not a picture of a hat. It was a picture of a boa constrictor digesting an elephant. But since the grown-ups were not able to understand it, I made another drawing: I drew the inside of the boa constrictor, so that the grown-ups could see it clearly.*

*Phrase 2. The grown-ups' response, this time, was to advise me to lay aside my drawings of boa constrictors, whether from the inside or the outside, and devote myself instead to geography, history, arithmetic and grammar. That is why, at the age of six, I gave up what might have been a magnificent career as a painter. I had been disheartened by the failure of my Drawing Number One and my Drawing Number Two. Grown-ups never understand anything by themselves, and it is tiresome for children to be always and forever explaining things to them. So then I chose another profession, and learned to pilot airplanes. I have flown a little over all parts of the world; and it is true that geography has been very useful to me. At a glance I can distinguish China from Arizona. If one gets lost in the night, such knowledge is valuable. In the course of this life I have had a great many encounters with a great many people who have been concerned with matters of consequence. I have lived a great deal among grown-ups. I have seen them intimately, close at hand. And that hasn't much improved my opinion of them. Whenever I met one of them who seemed to me at all clear-sighted, I tried the experiment of showing him my Drawing Number One, which I have always kept. I would try to find out, so, if this was a person of true understanding. But, whoever it was, he, or she, would always say: "That is a hat."*

**Português**

*Phrase 1.*

Claro que sim.

*Phrase 2.*

Meu desenho não representava um chapéu. Representava uma jibóia digerindo um elefante. Desenhei então o interior da jibóia, a fim de que as pessoas grandes pudessem compreender. Elas têm sempre necessidade de explicações.

*Phrase 3.*

*As pessoas grandes aconselharam-me deixar de lado os desenhos de jibóias abertas ou fechadas, e dedicar-me de preferência à geografia, à história, ao cálculo, à gramática. Foi assim que abandonei, aos seis anos, uma esplêndida carreira de pintor. Eu fora desencorajado pelo insucesso do meu desenho número 1 e do meu desenho número 2. As pessoas grandes não compreendem nada sozinhas, e é cansativo, para as crianças, estar toda hora explicando. 3 Tive pois de escolher uma outra profissão e aprendi a pilotar aviões. Voei, por assim dizer, por todo o mundo. E a geografia, é claro, me serviu muito. Sabia distinguir, num relance, a China e o Arizona. É muito útil, quando se está perdido na noite. Tive assim, no decorrer da vida, muitos contatos com muita gente séria. Vivi muito no meio das pessoas grandes. Vi-as muito de perto. Isso não melhorou, de modo algum, a minha antiga opinião. Quando encontrava uma que me parecia um pouco lúcida, fazia com ela a experiência do meu desenho número 1, que sempre conservei comigo. Eu queria saber se ela era verdadeiramente compreensiva. Mas respondia sempre: "É um chapéu". Então eu não lhe falava nem de jibóias, nem de florestas virgens, nem de estrelas. Punha-me ao seu alcance. Falava-lhe de bridge, de golfe, de política, de gravatas. E a pessoa grande ficava encantada de conhecer um homem tão razoável.*

**Spanish**

*Phrase 1.*

*Claro que si.*

*Phrase 2.*

Mi dibujo no representaba un sombrero. Representaba una serpiente boa que digiere un elefante. Dibujé entonces el interior de la serpiente boa a fin de que las personas mayores pudieran comprender. Siempre estas personas tienen necesidad de explicaciones.

*Phrase 3.*

Las personas mayores me aconsejaron abandonar el dibujo de serpientes boas, ya fueran abiertas o cerradas, y poner más interés en la geografía, la historia, el cálculo y la gramática. De esta manera a la edad de seis años abandoné una magnífica carrera de pintor. Había quedado desilusionado por el fracaso de mis dibujos número 1 y número 2. Las personas mayores nunca pueden comprender algo por sí solas y es muy aburrido para los niños tener que darles una y otra vez explicaciones. Tuve, pues, que elegir otro oficio y aprendía pilotear aviones. He volado un poco por todo el mundo y la geografía, en efecto, me ha servido de mucho; al primer vistazo podía distinguir perfectamente la China de Arizona. Esto es muy útil, sobre todo si se pierde uno durante la noche. 2 A lo largo de mi vida he tenido multitud de contactos con multitud de gente seria. Viví mucho con personas mayores y las he conocido muy de cerca; pero esto no ha mejorado demasiado mi opinión sobre ellas. Cuando me he encontrado con alguien que me parecía un poco lúcido, lo he sometido a la experiencia de mi dibujo número 1 que he conservado siempre. Quería saber si verdaderamente era un ser comprensivo. E invariablemente me contestaban siempre: "Es un sombrero". Me abstenía de hablarles de la serpiente boa, de la selva virgen y de las estrellas. Poniéndome a su altura, les hablaba del bridge, del golf, de política y de corbatas. Y mi interlocutor se quedaba muy contento de conocer a un hombre tan razonable.

With Aspell Dictionnaries

**Simple Analyser**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Phrase* | *Language* | *Matches found* | *Score (%)* | *TimeStamp result (ms)* |
| *1* | *English* | *1 (EN)* | *100* | *12* |
| *2* | *English* | *3 (EN, SP, PT)* | *89.39 (6.39, 4.22)* | *27* |
| *3* | *English* | *3 (EN, PT, SP)* | *90.56 (6.28, 3.16)* | *37* |
| *1* | *Portuguese* | *3 (PT, EN, SP)* | *52.63 (25.43, 21.93)* | *20* |
| *2* | *Portuguese* | *3 (PT, EN, SP)* | *77.88 (16.06, 6.06)* | *22* |
| *3* | *Portuguese* | *3 (PT, EN, SP)* | *78.20 (12.18, 9.62)* | *32* |
| *1* | *Spanhish* | *3 (PT, EN, SP)* | *(49.41) 36.62 (13.97)* | *18* |
| *2* | *Spanhish* | *3 (SP, EN, PT)* | *51.58 (34.23, 14.20)* | *19* |
| *3* | *Spanhish* | *3 (SP, EN, PT)* | *44.66 (33.57 21.77)* | *24* |

**Standard Analyser**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Phrase* | *Language* | *Matches found* | *Score (%)* | *TimeStamp result (ms)* |
| *1* | *English* | *1(EN)* | *100* | *12* |
| *2* | *English* | *3 (EN, SP, PT)* | *88.66 (6.82, 4.53)* | *25* |
| *3* | *English* | *3 (EN, PT, SP)* | *89.63 (6.91, 3.45)* | *33* |
| *1* | *Portuguese* | *3 (PT, SP, EN)* | *55.55 (22.60, 21.85)* | *20* |
| *2* | *Portuguese* | *3 (PT, EN, SP)* | *79.47 (14.45, 6.08)* | *20* |
| *3* | *Portuguese* | *3 (PT, EN, SP)* | *79.48 (10.93, 9.59)* | *23* |
| *1* | *Spanhish* | *3 (PT, EN, SP)* | *(50.58) 36.60 (12.82)* | *16* |
| *2* | *Spanhish* | *3 (SP, EN, PT)* | *52.62 (32.56, 14.82)* | *17* |
| *3* | *Spanhish* | *3 (SP, EN, PT)* | *45.96 (31.10, 22.94)* | *20* |

*This Analyzer uses the Tokenizer with StopWords, meaning that the articles are not considered.*

**Tests with N-Grams from 4 to 5**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Phrase* | *Language* | *Matches found* | *Score (%)* | *TimeStamp result (ms)* |
| *1* | *English* | *1 (EN, PT, SP)* | *83.73 (9.59, 6.67)* | *22* |
| *2* | *English* | *3 (EN, SP, PT)* | *72.81 (14.20, 13.00)* | *25* |
| *3* | *English* | *3 (EN, PT, SP)* | *76.67 (11.99, 11.34)* | *46* |
| *1* | *Portuguese* | *3 (SP, PT, EN)* | *(46.47) 37.52 (16.01)* | *17* |
| *2* | *Portuguese* | *3 (PT, SP, EN)* | *52.48 (23.85, 23.67)* | *22* |
| *3* | *Portuguese* | *3 (PT, SP, EN)* | *56.35 (25.20, 18.46)* | *34* |
| *1* | *Spanhish* | *3 (SP, PT, EN)* | *46.47 (37.52, 16.01* | *17* |
| *2* | *Spanhish* | *3 (SP, EN, PT)* | *45.32 (28.95, 25.73)* | *20* |
| *3* | *Spanhish* | *3 (SP, PT, EN)* | *47.62 (29.77, 22.61)* | *33* |

**Tests with N-Grams from 1 to 5**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Phrase* | *Language* | *Matches found* | *Score (%)* | *TimeStamp result (ms)* |
| *1* | *English* | *1 (EN, PT, SP)* |  |  |
| *2* | *English* | *3 (EN, SP, PT)* |  |  |
| *3* | *English* | *3 (EN, PT, SP)* |  |  |
| *1* | *Portuguese* | *3 (SP, PT, EN)* |  |  |
| *2* | *Portuguese* | *3 (PT, SP, EN)* |  |  |
| *3* | *Portuguese* | *3 (PT, SP, EN)* |  |  |
| *1* | *Spanhish* | *3 (SP, PT, EN)* |  |  |
| *2* | *Spanhish* | *3 (SP, EN, PT)* |  |  |
| *3* | *Spanhish* | *3 (SP, PT, EN)* |  |  |

**Tests with N-Grams from 4 to 10**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Phrase* | *Language* | *Matches found* | *Score (%)* | *TimeStamp result (ms)* |
| *1* | *English* | *1 (EN, SP, PT)* | *50.75 (26.84, 22.41)* | *26* |
| *2* | *English* | *3 (EN, PT, SP)* | *47.92 (26.71, 25.37)* | *40* |
| *3* | *English* | *3 (EN, PT, SP)* | *49.64 (25.44, 24.92)* | *62* |
| *1* | *Portuguese* | *3 (SP, PT, EN)* | *34.95 (33.87, 31.18)* | *22* |
| *2* | *Portuguese* | *3 (PT, SP, EN)* |  |  |
| *3* | *Portuguese* | *3 (PT, SP, EN)* |  |  |
| *1* | *Spanhish* | *3 (SP, PT, EN)* |  |  |
| *2* | *Spanhish* | *3 (SP, EN, PT)* |  |  |
| *3* | *Spanhish* | *3 (SP, PT, EN)* |  |  |