CENTRO DE INFORMÁTICA

**RELATÓRIO DA ATIVIDADE 1**

**MINERAÇÃO WEB**

Equipe:

Leonardo Schetini

Pedro Henrique

Ricardo Roster

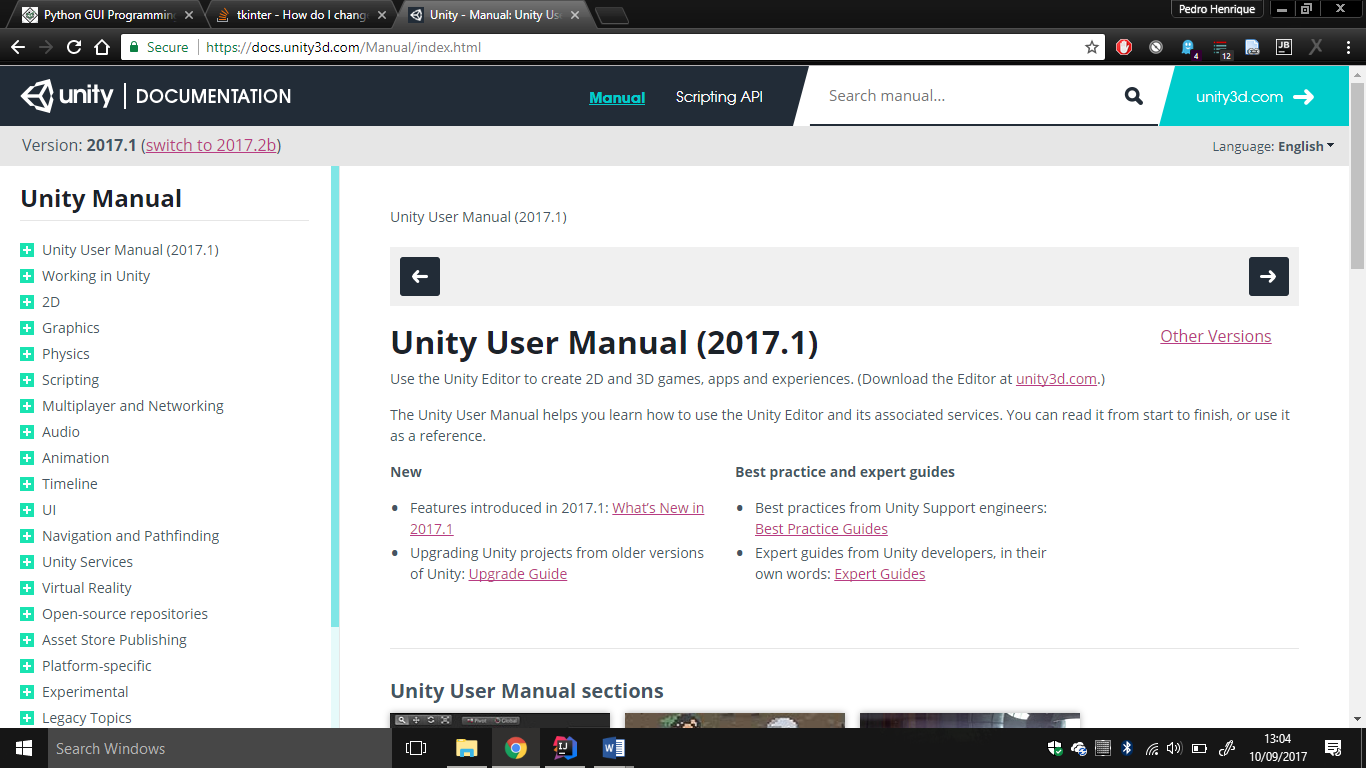
Recife - PE

Setembro / 2017

1. **CORPUS**

Os documentos selecionados são um conjunto de páginas que fazem parte da documentação da ferramenta Unity, especificamente o manual da ferramenta, consistindo de 1337 páginas HTML, mas como é necessário fazer a análise manual para duas consultas, foram selecionadas 300 páginas do total, informações como imagens, scripts e tags são removidas antes do processamento.

* 1. Exemplo de página



1. **AQUITETURA DO SISTEMA**

A ferramenta foi construída do zero usando a linguagem de programação python, usando apenas a lista de stopwords e o porter stemmer oferecidos pelo pacote nltk, e consiste de basicamente 3 módulos, são eles:

* text\_operators:
  + Esse módulo contém métodos para processamento de texto genérico, como remoção de stopwords e stemming, e também métodos para processar páginas html, removendo scripts e tags.
* inverted\_index
  + Módulo que constrói um índice invertido a partir do corpus, o índice guarda a informação da frequência dos termos em todos os documentos, frequência por documento e posições dos termos em cada documento.
* ranker
  + Cria um ranqueador para o corpus a partir do seu índice invertido, é basicamente o modelo de espaço vetorial, que usa TFIDF para construir os vetores dos documentos do corpus e das consultas e cosseno como medida de similaridade;
  + Ele é capaz de processar consultas com palavras e com contexto.

A ferramenta é usada pela aplicação, que consiste de um módulo que cria automaticamente quatro ranqueadores usando o corpus processado de maneiras diferentes, são elas:

* Clean - Processamento não remove stopwords ou aplica stemming;
* Stop - Processamento apenas remove stopwords;
* Stem - Processamento apenas aplica stemming;
* Stop\_stem - Processamento remove stopwords e stemming.

Em todas as configurações de processamento, os scripts e tags html são sempre removidos do corpus.

1. **CRIAÇÃO DE CONSULTAS E PREPARAÇÃO DOS TESTES**

O sistema é capaz de processar consultas com palavras isoladas e com contexto (parcialmente modificado dependendo do pré-processamento do corpus). As consultas usadas nos testes são:

* Creating **and** destroying objects;
* Mixing **and** playing audio.

Os documentos do corpus foram avaliados manualmente com valores fixos de 0 a 1 de acordo com um conjunto de regras definidas, são elas:

* 0.00 🡪 O documento não possui nenhum termo em como com a consulta;
* 0.10 🡪 O documento possui apenas um termo em comum com a consulta;
* 0.25 🡪 O documento possui mais de um termo em comum com a consulta, mas não todos;
* 0.50 🡪 O documento possui todos os termos em comum com a consulta, mas sua relevância para a mesma é baixa;
* 0.75 🡪 O documento possui todos os termos em comum com a consulta, sua relevância é razoável;
* 1.00 🡪 O documento possui todos os termos em comum com a consulta, sua relevância é alta.
  1. Resultado da Avaliação Manual
* 1ª linha 🡪 Documentos
* 2ª linha 🡪 Classificação manual de Creating **and** destroying objects
* 3ª linha 🡪 Classificação manual de Mixing **and** playing áudio

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Collider2D | CollidersOverview | CommandLineArguments | ComputeShaders | ComputingNormalPerpendicularVector | ConfiguringtheAvatar | Console | ContributingApply | ContributingBitbucket | ContributingClone | ContributingFAQ | ContributingFeedback | ContributingFork | ContributingPullRequest | ContributingReading | ContributingToUnity | ControllingGameObjectsComponents | ConventionalGameInput | Coroutines | CreateDestroyObjects | CreatingAndUsingScripts | CreatingComponents | CreatingGameplay | CreatingScenes | CrossPlatformConsiderations | CullingGroupAPI |
| 0,00 | 0,25 | 0,10 | 0,25 | 0,10 | 0,10 | 0,10 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,10 | 0,00 | 0,00 | 0,50 | 0,10 | 0,50 | 1,00 | 0,25 | 0,10 | 0,10 | 0,25 | 0,25 | 0,25 |
| 0,10 | 0,10 | 0,00 | 0,00 | 0,00 | 0,00 | 0,10 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,10 | 0,00 | 0,00 | 0,00 | 0,50 | 0,00 |
| Attributes | Audio | AudioFiles | AudioMixer | AudioOverview | AudioProfiler | AudioReference | AudioSpatializerSDK | AvatarCreationandSetup | BehindtheScenes | BestPracticeGuides | BlenderAndRigify | BlendShapes | BuildingAssetBundles4x | BuildPlayerPipeline | BuildSettings | Built-inShaderGuide | CacheServer | CameraRays | Cameras | CamerasOverview | CamerasReference | CameraTricks | CharacterControllers | ClusterRendering | CollaborateInProgress |
| 0,00 | 0,00 | 0,10 | 0,10 | 0,25 | 0,00 | 0,00 | 0,25 | 0,10 | 0,25 | 0,10 | 0,25 | 0,25 | 0,25 | 0,00 | 0,10 | 0,00 | 0,10 | 0,10 | 0,10 | 0,25 | 0,00 | 0,10 | 0,25 | 0,10 | 0,00 |
| 0,00 | 0,25 | 0,25 | 0,25 | 1,00 | 0,10 | 0,10 | 0,25 | 0,00 | 0,00 | 0,00 | 0,00 | 0,10 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |

1. **TESTES**

Nos testes é aplicado um limiar nos resultados classificados manualmente, qualquer documento com relevância maior que 0,2 é considerado relevante. O calculo dos valores de precisão, cobertura e f-measure, são feitos para 5, 10, 20, 50 e 100 documentos resultantes selecionados pelos ranqueadores.

* Consulta: Creating **and** destroying objects (92 documentos relevantes)
  + Resultados para 5 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 1 | 0,054348 | 0,103093 |
| Stop | 1 | 0,054348 | 0,103093 |
| Stem | 1 | 0,054348 | 0,103093 |
| Stop\_stem | 1 | 0,054348 | 0,103093 |

* + Resultados para 10 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,8 | 0,086957 | 0,156863 |
| Stop | 0,8 | 0,086957 | 0,156863 |
| Stem | 1 | 0,108696 | 0,196078 |
| Stop\_stem | 1 | 0,108696 | 0,196078 |

* + Resultados para 20 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,8 | 0,173913 | 0,285714 |
| Stop | 0,8 | 0,173913 | 0,285714 |
| Stem | 0,85 | 0,184783 | 0,303571 |
| Stop\_stem | 0,85 | 0,184783 | 0,303571 |

* + Resultados para 50 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,74 | 0,402174 | 0,521127 |
| Stop | 0,74 | 0,402174 | 0,521127 |
| Stem | 0,78 | 0,423913 | 0,549296 |
| Stop\_stem | 0,78 | 0,423913 | 0,549296 |

* + Resultados para 100 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,67 | 0,728261 | 0,697917 |
| Stop | 0,67 | 0,728261 | 0,697917 |
| Stem | 0,7 | 0,760870 | 0,729167 |
| Stop\_stem | 0,7 | 0,760870 | 0,729167 |

* Consulta: Mixing **and** playing audio (16 documentos relevantes)
  + Resultados para 5 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,8 | 0,25 | 0,380952 |
| Stop | 0,8 | 0,25 | 0,380952 |
| Stem | 0,6 | 0,1875 | 0,285714 |
| Stop\_stem | 0,6 | 0,1875 | 0,285714 |

* + Resultados para 10 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,6 | 0,375 | 0,461538 |
| Stop | 0,6 | 0,375 | 0,461538 |
| Stem | 0,5 | 0,3125 | 0,384615 |
| Stop\_stem | 0,5 | 0,3125 | 0,384615 |

* + Resultados para 20 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,5 | 0,625 | 0,555556 |
| Stop | 0,5 | 0,625 | 0,555556 |
| Stem | 0,5 | 0,625 | 0,555556 |
| Stop\_stem | 0,5 | 0,625 | 0,555556 |

* + Resultados para 50 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,28 | 0,875 | 0,424242 |
| Stop | 0,28 | 0,875 | 0,424242 |
| Stem | 0,3 | 0,9375 | 0,454545 |
| Stop\_stem | 0,3 | 0,9375 | 0,454545 |

* + Resultados para 100 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,15 | 0,9375 | 0,258621 |
| Stop | 0,14 | 0,875 | 0,241379 |
| Stem | 0,15 | 0,9375 | 0,258621 |
| Stop\_stem | 0,15 | 0,9375 | 0,258621 |

* Média das consultas
  + Resultados para 5 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,9 | 0,152174 | 0,2420225 |
| Stop | 0,9 | 0,152174 | 0,2420225 |
| Stem | 0,8 | 0,120924 | 0,1944035 |
| Stop\_stem | 0,8 | 0,120924 | 0,1944035 |

* + Resultados para 10 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,7 | 0,231 | 0,3092 |
| Stop | 0,7 | 0,231 | 0,3092 |
| Stem | 0,75 | 0,2106 | 0,2903 |
| Stop\_stem | 0,75 | 0,2106 | 0,2903 |

* + Resultados para 20 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,65 | 0,3995 | 0,4206 |
| Stop | 0,65 | 0,3995 | 0,4206 |
| Stem | 0,675 | 0,4049 | 0,4296 |
| Stop\_stem | 0,675 | 0,4049 | 0,4296 |

* + Resultados para 50 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,51 | 0,6386 | 0,4727 |
| Stop | 0,51 | 0,6386 | 0,4727 |
| Stem | 0,54 | 0,6807 | 0,5019 |
| Stop\_stem | 0,54 | 0,6807 | 0,5019 |

* + Resultados para 100 documentos:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisão | Cobertura | F-measure |
| Clean | 0,41 | 0,8329 | 0,4783 |
| Stop | 0,405 | 0,8016 | 0,4696 |
| Stem | 0,425 | 0,8492 | 0,4939 |
| Stop\_stem | 0,425 | 0,8492 | 0,4939 |

1. **CONCLUSÃO**

É possível notar uma diferença sutil nos resultados das bases, principalmente nas bases onde o stemming é aplicado, tendo estas uma média maior em relação as bases onde o stemming não é usado, exceto no caso onde são considerados apenas os 5 primeiros resultados, isso é causa da facilidade de encontrar mais resultados que contenham palavras parecidas com as da consulta, já que apenas o radical é usado. A remoção de stopwords praticamente não influencia o resultado, pois como é usado espaço de vetores com TFIDF e medida de similaridade do cosseno, os pesos desses termos são desprezíveis já que aparecem em muitos documentos.

Como apenas duas consultas foram usadas na validação, não é possível afirmar se a média com muitas consultas será parecida, pois até o resultado das duas consultas diferem bastante.