Universidade da Beira Interior
Departamento de Informática

Text Glow – Caracterização Estética e Qualitativa de um Texto

Elaborado por:

José Pedro Estima Santos

Orientador:

Professor Doutor João Paulo Cordeiro

8 de Junho de 2017

Agradecimentos

Gostaria de agradecer ao meu orientador, Professor Doutor João Paulo Cordeiro pela oportunidade de desenvolver este projeto porque expandiu o meu conhecimento em alguns campos. Quero agradecer também todo o suporte e motivação dados no âmbito deste projeto.

Gostaria também de agradecer a toda a minha família, aos meus pais, Luís Santos e Lurdes Estima, por me terem educado, apoiado e suportado ao longo deste percurso de todas as maneiras possíveis.

Eu quero também mostrar a minha gratidão a todos os professores com os quais tive a oportunidade de aprender, cada um me ensinou novas e diversas coisas, em especial toda a lógica e pensamento que envolve o desenho da soluções de problemas que aprendi ao longo deste percurso académico.

Finalmente, um grande abraço a todos os meus amigos que sempre me apoiaram e encorajaram.

Conteúdo

Lista de Figuras

Lista de Tabelas

List of Listings

Acrónimos

LD Lexical Diversity

TTR Type Token Ratio

VOC Vocabulário

ADJR Rácio de adjetivos

API Application Program Interface

MTLD Measure of Textual Lexical Diversity

VT Variance Time

RS Rescaled Range Statistics

KS Kolmogorov-Smirnov

Capítulo 1

Introdução

O Capítulo 1 é dedicado à apresentação do tema do projeto, assim como a justificação da sua escolha. Aqui serão ainda apresentados os seus objetivos bem como a organização deste documento.

Todos os dias grandes quantidades de texto são colocadas na Internet, tornando-se assim públicas e acessíveis a qualquer pessoa em qualquer parte do mundo. A maior parte destes textos (narrativas, postagens em blogues, notícias, entre outros) são, normalmente, redigidos por pessoas de diferentes idades, ocupações e níveis académicos.

Geralmente um leitor humano, com experiência suficiente, consegue avaliar a qualidade e complexidade de escrita de um determinado texto. Contudo. é facilmente percetível que as pessoas novas, nomeadamente crianças, não têm uma qualidade de escrita relativamente boa. Isto acontece devido ao seu vocabulário ainda não estar suficientemente desenvolvido para conseguirem interligar frases com os vários articuladores discursivos existentes ou embelezá-las com os mais variados recursos linguísticos. Por outro lado, uma pessoa adulta tem um melhor conhecimento da língua que está a utilizar. Portanto, teoricamente, a sua qualidade de escrita deve ser bastante superior à das crianças. Um adulto consegue usar uma grande variedade de articuladores de discurso perante as diferentes situações em que se encontra e, como tem acesso a um vocabulário mais abrangente e complexo, consegue facilmente fazer uso dos processos de coesão textual (sinonímia, hiperonímia, hiponímia, anáfora, catáfora, etc.) de modo a que o seu texto não se torne repetitivo e, consequentemente, pouco apelativo.

O facto de as pessoas, ao longo da sua vida, redigirem poucas produções escritas, mesmo na sua língua materna, bem como a falta de hábitos de leitura, origina uma mobilização deficiente do vocabulário. Já aqueles que fazem da construção de texto o seu modo de vida, os escritores, têm uma capacidade extraordinária de organizar uma história e embelezá-la com o uso de figuras de estilo. As figuras de estilo são uma espécie de artifícios estéticos de que os escritores se socorrem

não só na expressão de pensamentos ou imagens, como também nas construções sintáticas. Estas figuras pretendem, de uma forma geral, conferir ao texto originalidade, emotividade e, inclusive, teor poético. Por vezes, são empregues pelo autor com o objetivo de criar empatia entre o leitor e a personagem e/ou narrador. A título de exemplo desta identificação entre leitor e entidade textual, considere-se o texto poético "Autopsicografia" de Fernando Pessoa que se centra na relação que o "eu" cria com o seu próprio texto e na forma como essa relação é transmitida ao leitor. O "eu" constrói o seu texto partindo de uma emoção verdadeira que é, depois, filtrada pela subjetividade do "eu", resultando numa emoção fingida que é transmitida ao leitor. Ou seja, o leitor experiencia uma vivência que não é a verdadeira que o sujeito poético vivenciou, pois a criação literária exige, de acordo com a teoria que o "eu" pretende enunciar neste texto, um fingimento que é tido como o leitor como verdadeiro.

1.1 Motivação

Como foi mencionado na secção anterior, é notório que os escritores dão um toque único a um texto. Com base em estudos anteriores [1] [2], existem fortes indícios de que os padrões estéticos no texto poderão ser modelados por leis matemáticas. Apesar desta ser uma área muito recente, já existem alguns estudos sobre este assunto onde são comprovados os indícios subjacentes aos textos lexicalmente ricos.

O que ainda não existe é uma compilação de todas as medidas relevantes estudadas até ao momento numa ferramenta única que seja capaz de determinar efetivamente a riqueza lexical e beleza estética de um texto.

1.2 Objetivos

A qualidade e beleza de um texto depende de certos fatores. Embora alguns sejam já conhecidos, nomeadamente a riqueza e variedade léxica, o uso de determinadas características, as figuras de estilo, outros há que ainda são objetos de estudo e investigação científica.

Tudo indica que existem marcadores que podem ser usados para caracterizar a beleza e qualidade de escrita dos diferentes géneros literários. Estes marcadores permitem aos leitores terem uma perceção da estética que se encontra no texto de uma forma subliminar.

O objetivo deste projeto é estudar e experimentar certas medidas de caracterização estética de um texto, bem como a implementação de uma aplicação que mede de uma maneira muito aproximada a estética e beleza de um texto de forma

quantitativa.

1.3 Organização do documento

De modo a refletir o trabalho que foi feito, este documento encontra-se estruturado da seguinte forma:

1. O primeiro capítulo – Introdução – apresenta o projeto, a motivação para a sua escolha, o enquadramento para o mesmo, os seus objetivos e a respetiva organização do documento.

2. O segundo capítulo – Estado da Arte – descreve o que de mais relevante foi encontrado na literatura científica sobre o assunto da caraterização estética do texto.

3. O terceiro capítulo - Tecnologias e Ferramentas Utilizadas - descreve as tecnologias utilizadas durante do desenvolvimento do projeto.

4. O quarto capítulo - Resultados Obtidos - apresenta e discute os resultados obtidos através do uso da ferramenta desenvolvida.

5. O quinto capítulo - Conclusões e Trabalho Futuro - descreve as conclusões obtidas pelo autor, bem como algumas sugestões de desenvolvimento do projeto no futuro.

Capítulo 2

Estado da Arte

2.1 Introdução

O Capítulo 2 é dedicado ao que de mais relevante foi encontrado na literatura científica sobre o assunto da caraterização estética do texto [2] [1] .

2.2 Riqueza Léxica de um Texto

A riqueza ou diversidade léxica de um texto, Lexical Diversity (LD), estima a proporção entre palavras distintas (*types*) e ocorrências (*tokens*) no texto. Para entender isto, é fundamental entender os conceitos de *types* e *tokens*. À medida que um texto aumenta, o seu número de *tokens* aumenta também, ou seja, o número de *tokens* aumenta de forma linear com o tamanho do texto.

Apesar deste aumento ser linear, o aumento de *types* é mais lento, pois a adição de novos *tokens* não garante que sejam novos *types* devido à repetição de palavras ao longo de um texto. Isto acontece porque não existe nenhum texto com algumas dezenas de palavras que tenha sentido sem a repetição de algum tipo de *tokens*. Assim podemos afirmar que à medida que um texto fica mais longo existe um decréscimo no valor da diversidade que o representa. Um texto está totalmente representado em *types* quando atinge um ponto em que a partir deste não é possível encontrar novos *types* [2].

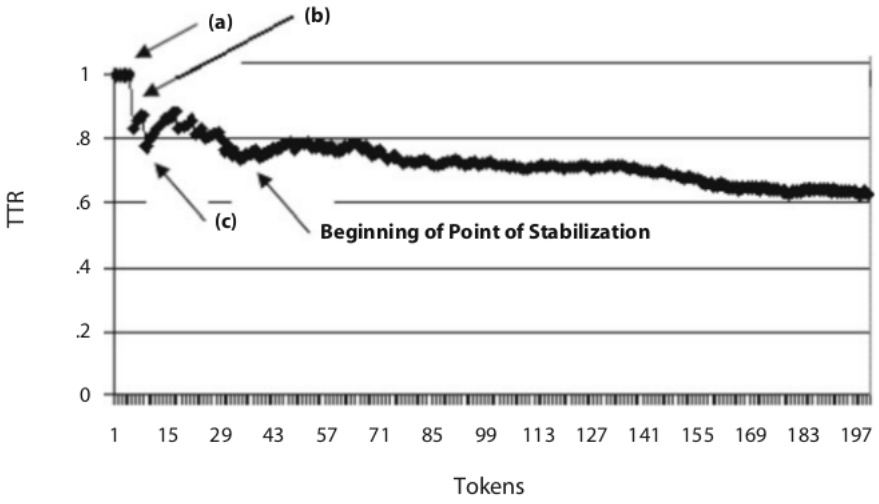

Figura 2.1: TTR e ponto de estabilização das primeiras 200 palavras do primeiro capítulo de "The Red Badge of Courage"retirada de [2].

Para perceber o ponto de estabilização temos de entender que inicialmente a sequência de texto que está a ser analisado apenas contém novos *types*. Na Figura 2.1, esta zona pode ser vista como a curva contínua (marcada com a letra a). Após encontrar o primeiro *type* repetido, o valor do TTR baixa drasticamente (marcado como b). A seguir, encontramos de novo uma sequência de novos *types*, isto resulta num aumento da curva do TTR, e novamente uma repetição de *types* causa outra baixa (marcado em c). Esta área de altos e baixos continua até se chegar ao ponto de estabilização. É neste ponto que nem a repetição de *types* ou um considerável aparecimento de novos *types* afetam a curva do TTR. Assim, podemos considerar que a zona da sequência após o ponto de estabilização é uma saturação, podemos ver esta manifestação na Figura 2.1 onde existe uma descida gradual e relativamente suave com que a curva acaba.

Todas as análises textuais apresentam muitas dificuldades e discordâncias e a LD não é diferente. Não existe, neste campo de estudo, uma concordância sobre qual a melhor forma de processar texto (sequencial ou não sequencial) ou de compor termos léxicos (lemas, palavras, bigramas). Também não é clara uma posição em relação às distinções entre os termos diversidade lexical, diversidade do vocabulário e riqueza léxica [2]. É óbvio que este campo de estudo é relativamente novo e ainda precisa de muita investigação para se chegar a conclusões mais sólidas. Portanto, neste projeto apenas estão incluídos os mais sofisticados índices de *LD* que estão atualmente disponíveis segundo McCarthy e Jarvis [2].

2.3 Estimativa com vocd-D

O *vocd-D* é uma medida usada para calcular LD. O cálculo do *vocd-D* é o resultado de uma série de amostras de palavras escolhidas aleatoriamente. Começa-se por retirar do texto 100 amostras de 35 palavras e calcular o seu TTR, guardando-se esse valor. Repete-se o procedimentos para amostras de 36 a 50 palavras. Após isso calcula-se a média de todos os TTR e obtém-se assim o valor do *vocd-D* [2].

```
double ttr = 0;
for (int i = 35; i <= 50; i++) {
    //calcular ttr de i palavras
    ttr = ttr + ttrOfSomeWords(oText.toString(), i);
}
Object.setVocd(String.valueOf(ttr / 16));
```

Listing 2.1: Exemplo de cálculo do vocd-D.

2.4 Estimativa com MTLD

O Measure of Textual Lexical Diversity (MTLD) é também uma medida usada para calcular LD. Consiste em percorrer o texto apenas uma vez e ir calculando o seu TTR. Quando este for menor que 0.720, o cálculo do TTR deverá ser reiniciado e uma outra variável *factors* que é iniciada a 0 deverá ser incrementada uma vez. Após isto, obtém-se o MTLD que é a divisão do número total de palavras num texto pelo número de *factors* calculados [2].

```
of (1.00) the (1.00) people (1.00) by (1.00) the (.800) people
    (.667) |||FACTORS = FACTORS + 1||| for (1.00) the (1.00)
    people (1.00) e assim sucessivamente.
```

Listing 2.2: Exemplo de cálculo do MTLD.

2.5 Fractalidade em Texto

Com base no artigo [1] é discutido se textos contém propriedades fractais, isto é, se vários atributos que caracterizam frases são auto-semelhantes. Para fazer isto, 7 textos foram analisados usando várias ferramentas estatísticas para determinar se as sequências empíricas para estes mesmos atributos são Gaussianas e auto-semelhantes. O teste de Kolmogorov-Smirnov (KS) e dois parâmetros de Hurst, Variance Time (VT) e Rescaled Range Statistics (RS), foram aplicados usando estimadores. Estes dois parâmetros definem o *Hurst Exponent* que é uma medida de longo prazo de séries temporais. Relaciona-se às autocorrelações da série temporal, e a velocidade a que elas diminuem.

Auto-semelhança é uma propriedade dos fractais e refere-se à possibilidade das partes de um objeto serem semelhantes ao todo. No caso particular de auto-semelhança, refere-se especificamente ao facto das propriedades estatísticas do objeto serem as mesmas independentemente da escala através da qual é observado.

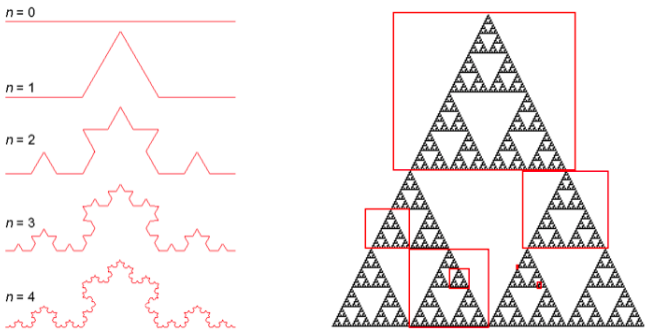

Figura 2.2: Exemplos de auto-semelhança em figuras geométricas.

Foi assim descoberto que existem dois contribuidores para a existência de auto-semelhança em textos. Primeiro, mostra-se que as sequências, construidas por vários atributos de um bloco de texto produzidos por humanos, são auto-semelhantes, sugerindo assim que o grau de auto-semelhança está relacionado com a qualidade do texto. Segundo, a maneira como a análise é feita define as fundações básicas para uma futura investigação na interseção destes dois campos, o campo da auto-semelhança e o campo da qualidade estética de um texto.

Os resultados mostram que existe, de facto, uma beleza fractal nos textos produzidos por humanos, naturalmente que esta propriedade não é incluída conscientemente pelos autores do texto, tornando assim os resultados apresentados muito mais interessantes. Resultados estes que sugerem que é possível que exista uma relação entre o que é entendido como a qualidade de um texto e o seu grau de auto-semelhança, contudo ainda é necessário uma análise mais completa e exaustiva.

Investigadores do Instituto de Física Nuclear da Polónia encontraram um padrão fractal complexo de frases na literatura, especificamente no texto *Finnegan's Wake* do autor *James Joyce* [5]. Foram feitas experiências com muitos textos, mas foi este que obteve os resultados mais promissores [4].

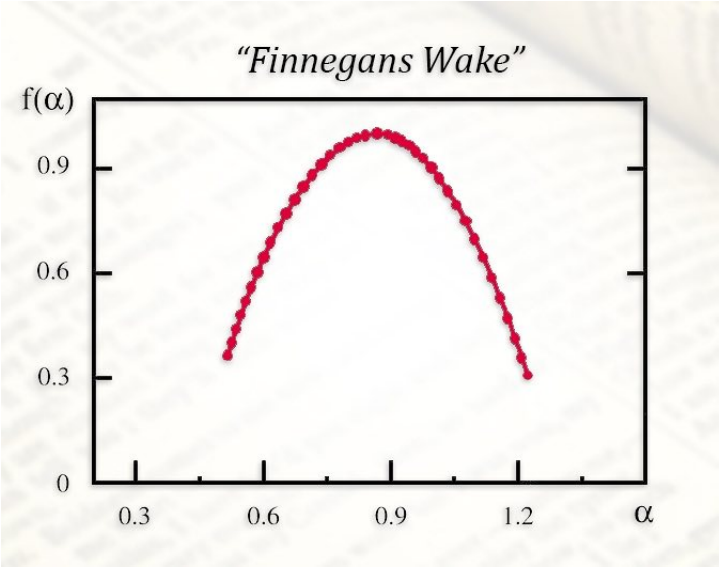

Figura 2.3: Análise multifractal de *Finnegan's Wake*, eixo horizontal representa o grau de silgularidade, eixo vertical representa o espetro de singularidade.

Capítulo 3

Tecnologias Utilizadas e Desenvolvidas

3.1 Introdução

O Capítulo 3 é dedicado à identificação e explicação das ferramentas e tecnologias usadas ao longo do desenvolvimento deste projeto. Foram usados os seguintes recursos, *JavaFX* para o desenvolvimento da interface gŕaica, a biblioteca Hultiglib, a biblioteca Apache POI para manipulação de texto em ficheiros *word* e *excel* e a biblioteca PDFBox para manipulação de textos em *PDF*.

3.2 Recursos Utilizados

3.2.1 JavaFX

JavaFX é uma plataforma que permite desenvolver aplicações com interface gráfica que usa como base programação por eventos. Nestas aplicações, os *eventos* são notificações de que algo aconteceu, como por exemplo, o utilizador carregou num botão ou numa tabela. Quando isto acontece, certos eventos são despoletados. Determinados filtros de eventos e *Event Handlers* dentro da aplicação recebem estes eventos e providenciam uma resposta [3].

Juntamente com o *SceneBuilder*, o *JavaFX* é uma excelente ferramenta para o desenvolvimento de aplicações com interface gráfica, estas duas ferramentas foram usadas durante o desenvolvimento deste projeto para construir a seguinte interface gráfica.

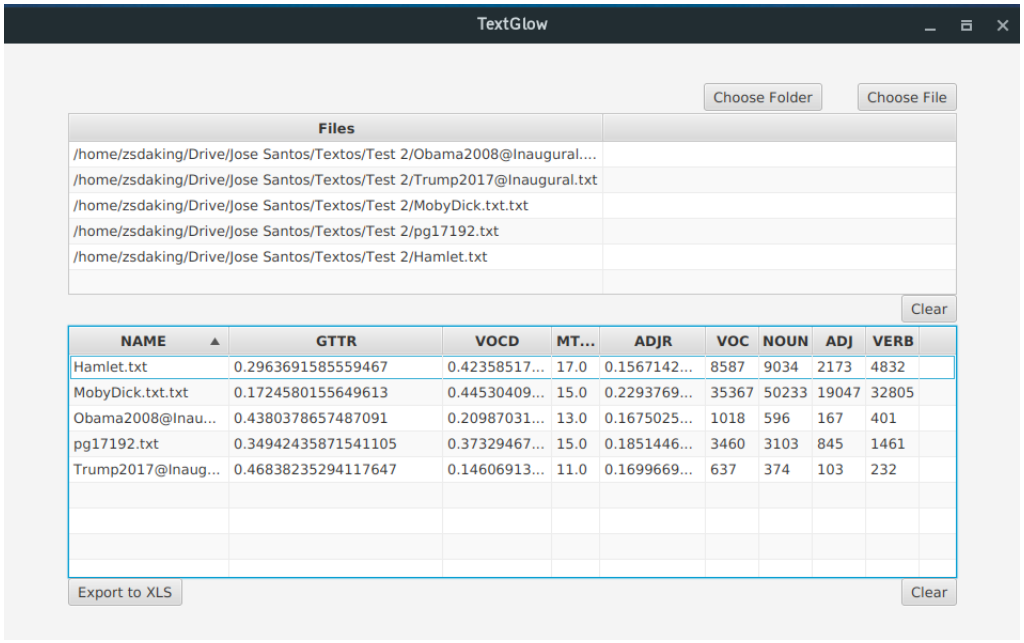

Figura 3.1: Interface gráfica do TextGlow.

Esta interface gráfica é muito simples de usar, tem a capacidade de analisar apenas texto a texto através da opção (*Choose File*), tem também a opção de analisar diretorias (*Choose Folder*), e são então analisados todos os textos dentro dessa pasta, possibilita a visualização de todos os textos e a diretoria onde se encontram que já foram analisados podendo assim apenas comparar os textos que quisermos, além da capacidade para exportar a tabela de resultados para um documento *.xls*, pode ser expandido para a adição de novas medidas.

3.2.2 Hultiglib

A biblioteca HultigLib tem sido desenvolvida no *Centro de Tecnologia da Linguagem Humana e Bioinformática* (HULTIG) da Universidade da Beira Interior, durante aproximadamente 10 anos. Esta biblioteca implementa, em linguagem *Java*, um vasto conjunto de funcionalidades para o *Processamento da Linguagem Natural* (PLN) que vão desde representação de entidades fundamentais (palavras, frases e textos) até à manipulação de relações e funções mais avançadas, como etiquetagem morfo-sintática, semelhança textual, deteção de paráfrases, entre outras.

No âmbito deste projeto foram usadas diversas funcionalidades da HultigLib, tais como instanciar objetos do tipo *Text*, bem como permitir a separação de um texto em frases, objetos do tipo *Sentece*, e em palavras, objetos do tipo *Word*. Contém ainda métodos que permitem fazer a etiquetagem morfossintática de todas as palavras contidas num texto, verificando se uma determinada palavra não é, de facto, uma palavra, (e.g. vírgulas, pontuação).

```
//Text
Text oText = new Text();
String s =
       "Pierre Vinken, 61 years old, will join the board as a"
      + "nonexecutive director Nov. 29. Mr. Vinken is chairman
         of "
      + "Elsevier N.V., the Dutch publishing group.";
oText.add(s);

//Sentence and Word
for (Sentence oSentence : oT) {
    oSentence = model.postag(oSentence);
    for (Word oWord : oSentence) {
        if (oWord.getTag().contains("JJ")) {
            iAdjCounter++;
        } else if (oWord.getTag().contains("VB")) {
                iVerbCounter++;
        } else if (oWord.getTag().contains("NN")) {
                iNounCounter++;
    }
}
```

Listing 3.1: Exemplo de uso de objetos da HultigLib.

Na Listing 3.1 podemos ver como são instanciados objetos palavra (Word), frase (Sentence) e texto (Text), além do método que nos permite conhecer a etiqueta sintática de uma palavra ((Word).getTag()). A HultigLib integra outras bibliotecas, com é o caso da *Apache OpenNLP* e do *Stanford Parser*, facilitando assim o uso de funcionalidades de processamento linguístico no processamento do texto.

3.2.3 Apache POI

Apache POI é uma Application Program Interface (API) em *Java* para processar documentos da *Microsoft*. Foi usada para ler texto de documentos *Word* e também para exportar as medidas calculadas para um documento *.xls*.

```
XWPFDocument docx = new XWPFDocument(new FileInputStream(path));
XWPFWordExtractor we = new XWPFWordExtractor(docx);
oText.add(we.getText().toLowerCase());
```

Listing 3.2: Retirar texto de um documento .docx.

```
HWPFDocument doc = new HWPFDocument(new FileInputStream(path));
WordExtractor we = new WordExtractor(doc);
oText.add(we.getText().toLowerCase());
```

Listing 3.3: Retirar texto de um documento .doc.

```
HSSFWorkbook workbook = new HSSFWorkbook();
HSSFSheet worksheet = workbook.createSheet("TextGlow Sheet");
```

Listing 3.4: Criar documento .xls.

```
HSSFRow row = worksheet.createRow((short) 0);
HSSFCell cell1 = row.createCell((short) 0);
cell1.setCellValue("NAME");
```

Listing 3.5: Criar uma linha e uma coluna num documento .xls.

```
workbook.write(fileOut);
```

Listing 3.6: Guardar ficheiro .xls.

3.2.4 PDFBox

PDFBox é uma biblioteca *open source* da *Apache* que permite a extração de texto de documentos *.pdf*. Foi usada para este mesmo fim.

```
try{
    PDDocument doc = PDDocument.load(new File(path));
    PDFTextStripper pdfStripper = new PDFTextStripper();
    oText.add(pdfStripper.getText(doc).toLowerCase());
}
```

Listing 3.7: Retirar texto de um documento pdf.

3.3 Recursos Desenvolvidos

3.3.1 ProjLib

A ProjLib é a biblioteca que desenvolvi no âmbito deste projeto. Nesta biblioteca encontram-se todos os métodos desenvolvidos durante este projeto para calcular LD. Nesta biblioteca encontram-se métodos como *getTextFromFile* que obtém o texto contido numa multiplicidade de formatos: *.txt*, *.pdf*, *.doc*, *.docx*.

O método *generalTTR* tem como objetivo calcular o TTR e Vocabulário (VOC) de todo o texto. O Vocabulário é o número de palavras diferentes num texto e o TTR consiste no rácio entre o número de palavras diferentes e o número total de palavras de um texto, conforme descrito na Secção 2.2.

O método *wordCaracterization* calcula o número de adjetivos, verbos e nomes num texto. Estes valores são depois usados no cálculo do ADJR, sendo que este é o rácio entre o número de adjetivos e a soma do número de verbos e nomes.

$$ADJR = \frac{\#ADJ}{\#NOUNS + \#VERBS}$$

O método *vocd-D* calcula o *vocd-D* e o método *mtLD* calcula o MTLD conforme referido no Capítulo 2.

3.3.2 ForC

Este programa calcula determinados atributos referidos no Capítulo 2, estes resultados são calculados com métodos da ProjLib convertendo todos os valores decimais para uma aproximação inteira. Estes resultados foram calculados para blocos de até 100 palavras. Os valores obtidos foram posteriormente submetidos à ferramenta *testH* (ferramenta utilizada para verificar fractalidade), os seus resultados irão ser discutidos no Capítulo 4.

3.4 Conclusões

Neste capítulo foram descritas todas as ferramentas utilizadas e produzidas no desenvolvimento deste projeto.

Capítulo 4

Resultados Obtidos

4.1 Introdução

O Capítulo 4 é dedicado à apresentação dos resultados que foram obtidos ao longo do desenvolvimento deste projeto, bem como a sua discussão. Como se sabe, existem vários tipos de géneros literários, literatura, romances, crónicas, fábulas, entre outros. No âmbito deste projeto, foram usados textos com diferentes características e diferentes estilos literários. Foram também usados diversos *posts* de *blogs* de autores de diversas faixas etárias, pelo que foram divididos em três grandes grupos, entre os 13-17 anos, 23-27 anos e 33-48 anos. Os textos utilizados foram os seguintes:

1. Discurso Inaugural da presidência dos Estados Unidos da América em 2017 [12]
2. The Raven [6]
3. Moby Dick [7]
4. Hamlet [8]
5. Alice In Wonderland [9]
6. Discuro Inaugural da presidêncida dos estados Unidos da América em 2008 [13]
7. Guilliver's Travels [10]
8. Millennium [11]

4.2 Apresentação e Discussão dos resultados obtidos

Aqui vão ser apresentadas todas as tabelas e valores obtidos para os vários textos que foram usados nos testes deste projeto. Considere-se a seguinte tabela:

#	Nome dos Textos	Nomes	Adjetivos	Verbos	Vocabulário
1	Trump2017@Inaugural	374	103	232	637
2	The Raven	3103	845	1461	3458
3	MobyDick	50233	19047	32805	35378
4	Hamlet	9034	2173	4832	8587
5	AliceInWonderland	1965	581	1985	2606
6	Obama2008@Inaugural	596	167	401	1018
7	Guilliver's Travels	21407	7476	16660	14213
8	Millennium	4062	1214	3714	5187

Tabela 4.1: Tabela de Nomes, Adjetivos, Verbos e Vocabulário.

Através da observação da Tabela 4.1 conseguimos ter referências, tais como os *nomes*, *adjetivos*, *verbos* e vocabulário, que nos permitem ter uma noção do tamanho do texto que foi usado. Estes valores são importantes pois permitem-nos criar a relação existente entre as medidas *vocd-D* e MTLD. Sendo assim, podemos facilmente constatar que os textos #1 e #6 são relativamente pequenos em relação aos outros, visto que esses textos são discursos inaugurais dos dois últimos presidentes dos EUA. Enquanto que os outros são obras literárias criadas por escritores de renome. Considere-se a seguinte tabela:

#	Nome dos Textos	General TTR	ADJR
1	Trump2017@Inaugural	0.4683824	0.1699670
2	The Raven	0.3492929	0.1851446
3	MobyDick	0.1725958	0.2293769
4	Hamlet	0.2963692	0.1567143
5	AliceInWonderland	0.2767924	0.1470886
6	Obama2008@Inaugural	0.4380379	0.1675025
7	Guilliver's Travels	0.1426192	0.1963906
8	Millennium	0.3032447	0.1561214

Tabela 4.2: Tabela de GTTR e ADJR.

Na Tabela 4.2 podemos observar que o rácio de adjetivos, ADJR, não reflete a qualidade de um texto, visto que os textos escritos por profissionais não têm um rácio muito maior que os textos dos discursos. Era esperado que o rácio de

adjetivos iria aumentar com o tamanho do texto, pois existindo mais texto existe também uma maior oportunidade dos escritores embelezarem mais o seu texto, contudo isso não se verifica.

Podemos também concluir que o General TTR não é uma boa medida para caracterizar a LD visto que, através da observação da tabela os valores dos textos supostamente "mais pobres" têm valores muito mais altos do que os textos escritos por profissionais e escritores de renome. Vejamos agora o efeito da medição com *vocd-D* e MTLD apresentado na Tabela 4.3.

#	Nome dos Textos	*vocd-D*	MTLD	*vocd-ADJR*
1	Trump2017@Inaugural	0.1463260	11	0.1521772
2	The Raven	0.3727139	15	0.1355566
3	MobyDick	0.4456147	15	0.1528902
4	Hamlet	0.4233622	17	0.1149562
5	AliceInWonderland	0.3050539	15	0.1062153
6	Obama2008@Inaugural	0.2102552	13	0.1253743
7	Guilliver's Travels	0.3778666	15	0.1400001
8	Millennium	0.3858613	14	0.1198797

Tabela 4.3: Tabela de vocd-D, MTLD e vocd-ADJR.

Ao observarmos a Tabela 4.3 podemos concluir duas coisas. Inicialmente conseguimos verificar que a medida *vocd-D* é uma boa medida para determinar indices de LD. Os textos que são "mais pobres" lexicalmente têm valores muitos baixos, o que já era esperado, enquanto que obras melhor conceituadas como a *Moby Dick* e o *Hamlet*, que são textos conhecidos universalmente por boas razões têm valores altos, alguns destes textos chegam a ter praticamente mais do dobro dos tais textos considerados "mais pobres" lexicalmente.

Com isto é possível verificar que ambas as medidas fazem distinção entre a qualidade associada ao texto, contudo o MTLD por ser efetuado com unidades inteiras não tem uma discriminação tão bom e distintiva da qualidade de um texto como o *vocd-D*. No MTLD conseguimos igualmente distinguir os textos de boa qualidade daqueles de fraca qualidade, contudo esta diferença pode não ser bem percebida com esta medida pois os valores são relativamente próximos uns dos outros. Com a medida *vocd-D* conseguimos notar uma maior distinção sendo que existem textos que têm o dobro do índice de qualidade de outros, enquanto que no MTLD esta situação não é tão fácil de verificar.

A medida *vocd-ADJR* foi posteriormente pensada e envolve o mesmo processo da *vocd-D*, em vez do TTR foram usados os valores necessários para calcular o ADJR, novamente foram escolhidas palavras aleatórias, 100 amostras de 35 a 50 palavras, e para cada um desses grupos foram efetuados os cálculos necessários

para determinar o ADJR de cada um destes grupos e por fim calcular uma média. Os resultados obtidos não devem ser usados para medir a qualidade de um texto, pois têm grandes discrepâncias em relação às outras medidas e sendo assim os resultados não são confiáveis, existem textos com boa qualidade, tanto nas medidas *vocd-D* e MTLD, que não obtêm uma diferenciação percetível em relação a outros, como por exemplo o texto #3 do #1.

Como foi dito anteriormente, foram feitos alguns testes sobre a fractalidade em textos de blogs escritos por pessoas de várias faixas etárias que foram posteriormente divididas em 3 grandes grupos. Considere agora a seguinte tabela:

Blogs - Idade	**VT**	**RS**	**KS**
Blogs - 13-17	0.692909	0.626349	0.698568
Blogs - 23-27	0.810752	0.563609	0.854745
Blogs - 33-48	0.608693	0.632027	0.539403

Tabela 4.4: Tabela obtida através da execução da ferramenta *testH* em blogs de várias idades para o ADJR.

Para facilitar a perceção destes valores foram atribuídos números aos vários grupos das faixas etárias, sendo que 13-17 é o grupo 1, 23-27 é o grupo 2 e por fim 33-48 é o grupo 3. Com a observação da Tabela 4.4 podemos verificar que o VT difere entre estes 3 grupos, o grupo 2 é o que apresenta um VT mais elevado enquanto que o grupo 3 apresenta o VT mais baixo e o grupo 1 se situa no meio. Em relação ao RS podemos verificar que as coisas já alteraram em relação ao VT, sendo que o RS mais alto pertence ao grupo 3, o mais baixo ao grupo 2 e novamente o grupo 1 apesar de se situar novamente no meio destes dois grupos está bastante próximo do grupo 3. Em relação ao KS verifica-se praticamente o mesmo que no VT, o grupo 2 é o que apresenta valores mais elevados seguidos do grupo 1 e por fim do grupo 3. Através da análise destes valores podemos concluir que o ADJR não é uma distribuição normal, pois o valor de KS é maior que 0.05, logo nada se pode concluir em relação à auto-semelhança.

Também foi analisado o TTR em termos de fractalidade, considere então a seguinte tabela: [1]

[1] VFE - Valor fora de escalar

Blogs - Idade	VT	RS	KS
Blogs - 13-17	0.69288	0.652863	0.000124
Blogs - 23-27	VFE	0.756867	0.000001
Blogs - 33-48	VFE	0.675235	0.000005

Tabela 4.5: Tabela obtida através da execução da ferramenta *testH* em blogs de várias idades para o TTR.

Mais uma vez para facilitar a perceção destes valores os grupos mantêm-se os mesmos. Através da observação da Tabela 4.5 nada se pode concluir para os valores de VT visto que os valores do grupo 2 e 3 são desconhecidos. Quanto ao RS o grupo 2 é aquele que apresenta o valor mais alto, seguido pelo grupo 3 e por fim o grupo 1. Em relação ao KS todos os grupos apresentam valores extremamente baixos sendo que o grupo 1 é o menos baixo, seguido do grupo 3 e por fim o grupo 2. Através da análise destes valores podemos concluir que o TTR é uma distribuição normal, pois o valor de KS é menor que 0.05, e portanto os valores VT e RS traduzem auto-semelhança em todos os grupos. Deste modo podemos verificar que em termos de TTR o grupo 2 é o mais auto-semelhante.

Capítulo 5

Conclusões e Trabalho Futuro

Neste capítulo são apresentadas as conclusões do autor sobre o trabalho desenvolvido.

5.1 Conclusões principais

Pode-se assim concluir, que existem, de facto princípios matemáticos por trás da qualidade de escrita de um texto. Através das experiências realisadas com as diversas medidas consegue-se concluir que, o *GeneralTTR* não é uma boa medida pois depende do tamanho do texto, textos pequenos têm claramente um *GeneralTTR* alto pois existem poucas repetições de palavras, enquanto que para os bons textos é inevitável não repetir palavras.

O *vocd-D*, é uma boa medida para obter a qualidade de um texto pois consegue identificar textos ricos lexicalmente de textos mais pobres, consegue-se ainda obter uma discriminação bem percetível dessa mesma qualidade.

O MTLD também é uma boa medida, consegue-se distinguir novamente os textos ricos lexicalmente dos pobres, contudo, como é uma medida unitária torna a discriminação dos textos uma tarefa difícil pois os valores relativamente altos são todos muito próximos uns dos outros, enquanto que no *vocd-D* conseguimos uma distinção melhor entre os textos #3 e #5 no MTLD essa distinção não é percetível.

O *vocd-ADJR* não é uma boa medida, visto que não existe praticamente distinção entre os textos #1 e #3, quando claramente a diferença de qualidade de uma obra literária para um discurso presidencial é enorme, algo que não se verifica neste caso.

Em relação à fratalidade, como a distribuição da medida ADJR não é normal, devido aos valores de KS maiores que 0.05 para todos os grupos, nada se pode concluir relativamente à auto-semelhança da mesma. Quanto à medida TTR, apresenta uma distribuição normal para todos os grupos, pois o valor de KS é menor

que 0.05, portanto existe auto-semelhança.

5.2 Trabalho Futuro

Grande parte do trabalho futuro incidirá sobre o estudo e desenvolvimento de novas medidas que suportem a existência de princípios matemáticos no que diz respeito à qualidade estética de um texto. Também passará por melhorar as medidas já existentes, algumas delas são muito pesadas computacionalmente com um grande grau de complexidade, idealmente era bom conseguir usar estas mesmas medidas com um grau de complexidade bastante inferior, logarítmico, poupando assim recursos e aumentando a velocidade de processamento da informação.

Bibliografia

[1] João Cordeiro, Pedro R. M. Inácio, and Diogo A. B. Fernandes. *Fractal Beauty in Text*. Portuguese Conference on Artificial Intelligence. Springer International Publishing. (2015): 796-801.

[2] Philip M. McCarthy and Scott Jarvis. *MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment*. Behavior research methods 42.2 (2010): 381-392.

[3] *https://docs.oracle.com/javafx/2/events/jfxpub-events.htm*

[4] *https://www.theguardian.com/books/2016/jan/27/scientists-reveal-multifractal-structure-of-finnegans-wake-james-joyce*

[5] Joyce, J., 2015. *Finnegans wake*. Penguin UK.

[6] Poe, E.A., Baudelaire, C. and Prüssen, E., 2000. *The Raven. Infomotions*, Incorporated.

[7] Melville, H., 1983. Redburn: His First Voyage; White-Jacket or the World in a Man-of-War; *Moby Dick or, The Whale*. Library of America.

[8] Shakespeare, W., Olivier, L. and Simmons, J., 1948. *Hamlet*. University Press.

[9] Carroll, L. and Pertwee, J., 1947. *Alice in wonderland*. Children's Press.

[10] Swift, J., 1995. *In Gulliver's Travels* (pp. 27-266). Palgrave Macmillan US.

[11] Everett, B. Cole, 2011. *Millennium*. Aegypan. ISBN:978-1463896423

[12] https://www.whitehouse.gov/inaugural-address

[13] https://obamawhitehouse.archives.gov/the-press-office/2013/01/21/inaugural-address-president-barack-obama