Uma aplicação para a ordenação do texto em função da complexidade oracional

R para o processamento de textos

Uma aplicação para a ordenação do texto em função da complexidade oracional

- Introdução
- R para o processamento de textos
- O processamento de orações
- Resultados
- Trabalho futuro e conclusões

Introdução: Trabalho prévio

- Processamento de texto para trabalho de corpus com PHP e XAMPP (linguística de corpus) GitHub - afonsoxavier/corpus
- Trabalho com R com fins pedagógicos (cursos de Língua e Sociedade com grande tradição quantitativa e de análise de dados): Language and Society (Spring 2015) | Afonso Xavier Canosa - Academia.edu
- PLN e trabalho com R com fins experimentais (experimentos de linguística, tese de doutoramento sobre geoentidades). R como principal linguagem de programação. CiTIUS | A identificacão e referenciacão de entidades geográficas mencionadas: o caso da 'Peregrinação', de Fernão Mendes Pinto

O processamento de textos em função da complexidade oracional

- Texto que queremos atacar com uma ordem diferente à linha narrativa. Fundamentalmente porque o nosso interesse não de leitura convencional senão de trabalho com o texto:
- Corpus dourado (parsing)
- Elaboração de gramáticas
- Alinhamento
- Treino para aprendizado de máquina

Materiais

- Texto: corpus suficientemente relevante como para cumprir a primeira lei de Zipf

- Ambiente de programação R. Livrarias para o processamento de texto (stringr, TM, tokenizers)

Procedimento (Ex. Texto 1)

Do que passey em minha mocidade neste Reyno ate que me embarquey para a India.Qvando às vezes ponho diante dos olhos os muitos e grãdes trabalhos e infortunios que por mim passarão, começados no principio da minha primeira idade, e continuados pella mayor parte, e milhor tẽpo da minha vida, acho que com muita razão me posso queixar da vẽtura que parece que tomou por particular tenção e empreza sua perseguirme, e maltratarme, como se isso lhe ouuera de ser materia de grande nome, e de grande gloria, porque vejo que não contente de me por na minha patria logo no começo da minha mocidade, em tal estado que nella viui sempre em miserias, e em pobreza, e não sem alguns sobresaltos e perigos da vida me quis tambẽ leuar às partes da India, onde em lugar do remedio que eu hia buscar a ellas, me forão crecendo com a idade os trabalhos, e os perigos. Mas por outra parte quãdo vejo que do meyo de todos estes perigos e trabalhos me quis Deos tirar sempre em saluo, e porme em seguro, acho que não tenho tanta razão de me queixar por todos os males passados, quãta de lhe dar graças por este só bẽ presente, pois me quis conseruar a vida, paraque eu pudesse fazer esta rude e tosca escritura, que por erança deixo a meus filhos (porque só para elles he minha tenção escreuella) paraque elles vejão nella estes meus trabalhos, e perigos da vida que passei no discurso de vinte e hũ ãnos em que fuy treze vezes catiuo, e dezasete vendido, nas partes da India, Etiopia, Arabia felix, China, Tartaria, Macassar, Samatra, e outras muitas prouincias daquelle oriental arcipelago, dos confins da Asia, a que os escritores Chins, Siames, Gueos, Elequios nomeão nas suas geografias por pestana do mũdo, como ao diante espero tratar muito particular...

Procedimento (Ex. Texto 2)

Pwyll Pendeuic Dyuet a oed yn arglwyd ar seith cantref Dyuet. A threigylgweith yd oed yn Arberth, prif lys idaw, a dyuot yn y uryt ac yn y uedwl uynet y hela. Sef kyueir o'y gyuoeth a uynnei y hela, Glynn Cuch. Ac ef a gychwynnwys y nos honno o Arberth, ac a doeth hyt ym Penn Llwyn Diarwya, ac yno y bu y nos honno. A thrannoeth yn ieuengtit y dyd kyuodi a oruc, a dyuot y Lynn Cuch i ellwng e gwn dan y coet. A chanu y gorn a dechreu dygyuor yr hela, a cherdet yn ol y cwn, ac ymgolli a'y gydymdeithon. Ac ual y byd yn ymwarandaw a llef yr erchwys, ef a glywei llef erchwys arall, ac nit oedynt unllef, a hynny yn dyuot yn erbyn y erchwys ef. Ac ef a welei lannerch yn y coet o uaes guastat; ac ual yd oed y erchwys ef yn ymgael ac ystlys y llannerch, ef a welei carw o ulaen yr erchwys arall. A pharth a pherued y llannerch, llyma yr erchwys a oed yn y ol yn ymordiwes ac ef, ac yn y uwrw y'r llawr.

Procedimento

- O texto foi depurado de marcas de anotação (tags tipo XML), ficando um documento com apenas as palavras e sinais de pontuação.
- Os tokens são então agrupados em tipos únicos e reordenados numa distribuição zipfiana.

List of terms

Pwyll Pendefig Dyfed: Peniarth 4 (Llyfr Gwyn Rhydderch) (t1r c1 l1 - t10r c38 l11)

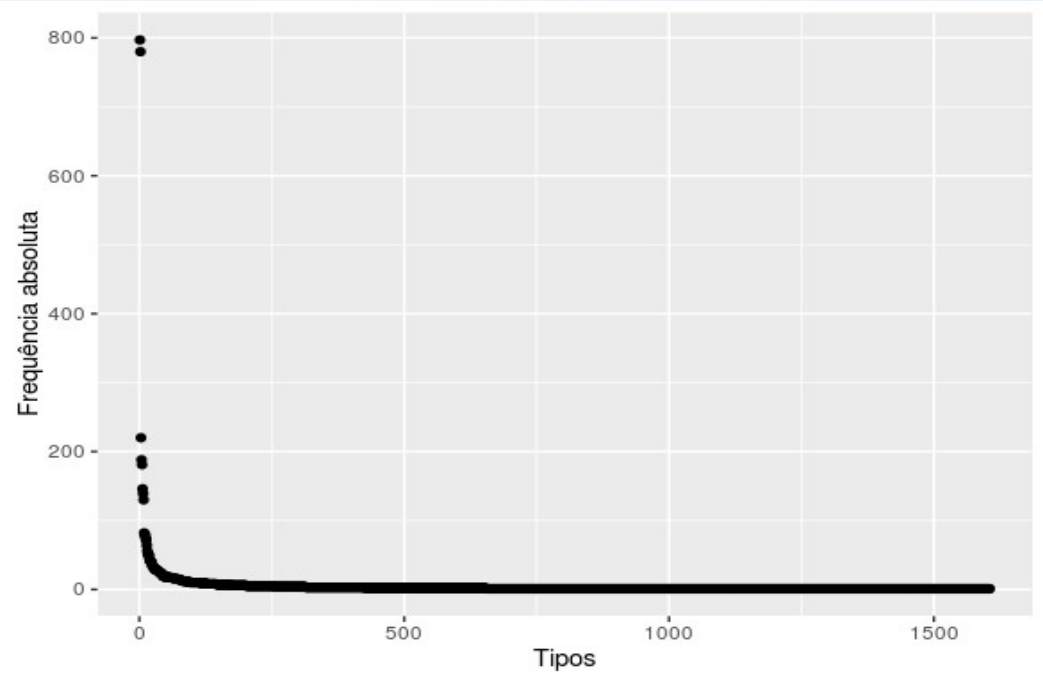
A total of **7846** word units have been processed.

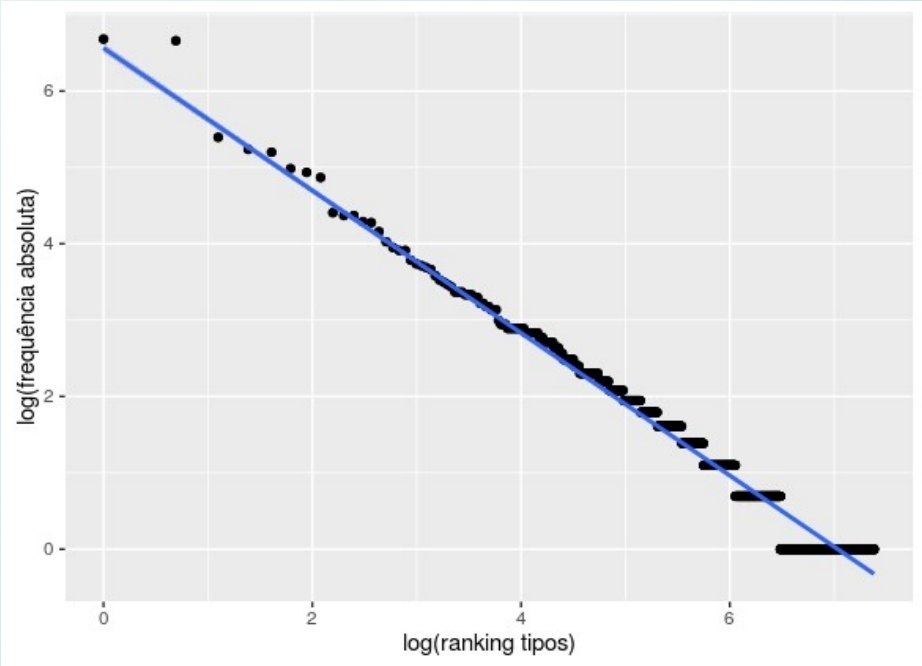
A total of **1605** unique terms have been extracted.

term 1 = **y** has **797** occurrences, **10.16**% of target text.
term 2 = **a** has **780** occurrences, **9.941**% of target text.
term 3 = **ac** has **220** occurrences, **2.804**% of target text.
term 4 = **heb** has **188** occurrences, **2.396**% of target text.
term 5 = **yn** has **181** occurrences, **2.307**% of target text.
term 6 = **ef** has **146** occurrences, **1.861**% of target text.
term 7 = **r** has **139** occurrences, **1.772**% of target text.
term 8 = **o** has **130** occurrences, **1.657**% of target text.
term 9 = **ar** has **82** occurrences, **1.045**% of target text.
term 10 = **i** has **79** occurrences, **1.007**% of target text.
term 11 = **yr** has **79** occurrences, **1.007**% of target text.

Procedimento

- Os tokens são então agrupados em tipos únicos e reordenados numa distribuição zipfiana.





A reordenação da oração

- É possível capturar a complexidade oracional?
- Reordenar o texto em função de preferências que atendam a possível dificuldade do texto para a sua compreensão sintática e semântica?

A reordenação da oração

- É possível capturar a complexidade oracional?
- Premisas:
- Maior longitude da oração implica maior número de unidades e relações sintáticas.
- Maior frequência dum termo implica maior conhecimento semântico (ou mais contexto para capturá-lo).

A reordenação do texto em função do valor da oração

- Maior longitude da oração (**t**) implica maior número de unidades e relações sintáticas.
- Maior frequência (**f**) de um termo implica maior conhecimento semântico e mais contexto para capturá-lo no conjunto do corpus (**T**).

Então o valor da oração O = $$\frac{\left(\sum_{i=1}^{t} \frac{f_i}{T}\right)^{w}}{t^{s}}$$

Onde **s** e **w** definem as preferências do usuário para um maior peso da sintaxe ou da semântica.

Resultados

Orações no topo da ordenação

A) Frequência dos termos com a maior relevância

- **1.** A chyuodi y uynyd, a dodi y deudroet yn y got, a troi o Pwyll y got yny uyd Guawl dros y penn yn y got ac yn gyflym caeu y got, a llad clwm ar y carryeu, a dodi llef ar y gorn.
  2. Ef a aeth ryngtaw a llys Eueyd Hen, ac ef a doeth y r llys, a llawen uuwyt wrthaw, a dygyuor a llewenyd ac arlwy mawr a oed yn y erbyn, a holl uaranned y llys wrth y gynghor ef y treulwyt.
  3. Ac yskynuaen a oed odieithyr y porth, eisted gyr llaw hwnnw beunyd, a dywedut y pawb a delei o r a debygei nas gwyppei, y gyffranc oll, ac o r a attei idi y dwyn, kynnic y westei a phellynic y dwyn ar y cheuyn y r llys.
  4. Ef a gyuodes Pwyll y uynyd, a pheri dodi gostec, y erchi y holl eircheit a cherdoryon dangos, a menegi udunt y llonydit pawb o honunt wrth y uod a y uympwy ; a hynny a wnaethpwyd.
- **B) Frequência dos termos mais relevante que a complexidade da oração**
- **1.**A hynny a wnaeth y makwyf.
  2.Y llys a gyrchyssant.
  3.A chyrch y llys.
  4.Y vely a gyrchwys, a y vreic a aeth attaw.

$$\frac{\left(\sum_{i=1}^{t} \frac{f_i}{T}\right)^{w}}{t^{s}}$$

Resultados

Orações no topo da ordenação

- **C) Complexidade da oração mais relevante que a frequência dos termos**
- **1.** heb y Pwyll.
  2. Y llys a gyrchyssant.
  3.A chyrch y llys.
  4. Ni a adwa.
- **D) Complexidade da oração muito relevante, frequência dos termos relevante**
- **1.** heb ef.
  2. heb y Pwyll.
  3. heb hi.
  4. Ni a adwa.
- **E) Complexidade da oração com a maior relevância**
- **1.** heb ef.
  2. heb hi.
  3. heb wy.
  4. Paham?

$$\frac{\left(\sum_{i=1}^{t} \frac{f_i}{T}\right)^w}{t^s}$$

Resultados (Exemplo 2)

$$\frac{\left(\sum_{i=1}^{t} \frac{f_i}{T}\right)^{w}}{t^{s}}$$

> print (wf_more[1:5,1])

- [1] De maneyra que primeyro que o Visorrey se tornasse a refazer do que perdera, e ajuntar o que a tormenta lhe espalhara por diuersas partes, se passou mais de hum mès

- [2] E abraçandose comigo muyto apertadamente, me pidio com muytas lagrimas que logo o fizesse Christão, porque entendia, e assi o confessaua que só com o ser se podia saluar, e não na triste seita de Mafamede, em que ate então viuera, de que pedia a Deos perdão

- [3] Da armada que o Achem mandou contra el Rey de Aarû, e do que lhe socedeo chegando ao rio de Paneticão

- [4] Outros dizião que era o Patemarcaa, com as cem fustas do Çamorim Rey de Calecù, outros todauia dizião que erão Turcos, e assi o affirmauão por rezoẽs muyto claras e euidentes
- [5] Como este Rey Bata partio de Turbão para o Achem, e do que fez despois que se rio com elles

Resultados (Exemplo 2)

print("Sentence complexity more relevant, word frequency highly relevant”)

print (sent_more[1:5,1])

$$\frac{\left(\sum_{i=1}^{t} \frac{f_i}{T}\right)^{w}}{t^{s}}$$

- [1] O Senhor que nos criou nos defenderá
- [2] E com isto me torno a meu proposito
- [3] Do que me aconteceo despois que me party deste reyno de Aarû
- [4] E dos Aarùs morreraõ sós quatrocentos
- [5] Do que mais me socedeo com este mercador Mouro

Resultados (Exemplo 2)

> print("Sentence complexity more relevant")

> print (sent_wf[1:5,1])

- [1] O Senhor que nos criou nos defenderá
- [2] E com isto me torno a meu proposito
- [3] E dos Aarùs morreraõ sós quatrocentos
- [4] Do que mais me socedeo com este mercador Mouro
- [5] De Panajù, aos cinco mamocos da oitaua Lũa

$$\frac{\left(\sum_{i=1}^{t} \frac{f_i}{T}\right)^w}{t^s}$$

Resultados (Exemplo 2)

> print("Sentence complexity more relevant")

> print (sent_wf[1:5,1])

- [1] O Senhor que nos criou nos defenderá
- [2] E com isto me torno a meu proposito
- [3] E dos Aarùs morreraõ sós quatrocentos
- [4] Do que mais me socedeo com este mercador Mouro
- [5] De Panajù, aos cinco mamocos da oitaua Lũa

$$\frac{\left(\sum_{i=1}^{t} \frac{f_i}{T}\right)^w}{t^s}$$

Conclusões e trabalho futuro

- A reordenação do texto utilizando apenas variáveis quantitativas oferece resultados significativos atendendo a critérios sintáticos e semânticos.
- Os valores *w* e *s* modulam a relevância da sintaxe ou do vocabulário. São valores suxeitos a preferências do/a usuário/a que permitem obter as diferentes ordenações. É no cálculo destes valores que se situa uma das linhas de trabalho futuro. Cálculo a partir de treino com grandes corpora? Aprendizado de máquina?

Conclusões e trabalho futuro

- Aplicações (já agora)
- Criação de corpora para a aprendizagem de modelos de língua para PLN.
- Aceleração dos resultados base na otimização de sistemas para córpora reduzidos (ex. históricos, variantes não estándar)

Scripts e corpus para replicar o experimento:

https://github.com/afonsoxavier/order_sentences

Obrigado!

Bibliografia

[1] Wickham, H. (2022). *stringr: Simple, Consistent Wrappers for Common String Operations*. http://stringr.tidyverse.org, https://github.com/tidyverse/stringr.
[2] Mullen, L., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, Consistent Tokenization of Natural Language Text. *Journal of Open Source Software*, *3*(23), 655.
[3] Li, W. (1991). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on information theory*, *38*(6), 1842-1845
[4] Piantadosi, S. T. (2014). Zipf’s word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112-1130.
[5] Kornai, A. (2008). *Mathematical Linguistics*. London: Springer.