**Passo 4**

Após a limpeza, procuramos criar um modelo estatístico para identificar frequências anômalas entre semestres, em especial, estamos interessados no caso em que a frequência em que um termo aparece em um semestre cresce rapidamente no semestre seguinte. Como modelo usamos o processo de Poisson, que é um modelo matemático para eventos aleatórios habitualmente utilizado em teoria das filas.

Um processo de Poisson com frequência *λ* no intervalo [0, ∞) é um gerador aleatório em que:

1. O número de eventos que ocorrem em um período *t* é uma variável aleatória com distribuição de Poisson (*λ)*.
2. O número de eventos em intervalos disjuntos (sem sobreposição) são variáveis aleatórias independentes.

Geralmente, processos de Poisson são utilizados para, dado o conhecimento de uma certa frequência *λ* em um período t no passado, estimar a probabilidade de o mesmo evento ocorrer um número *k* de vezes em um período *t+1* no futuro. Este, entretanto, não é o nosso caso. Como dito anteriormente, não estamos fazendo nenhum tipo de predição sobre o potencial de termos se tornarem bastante mencionados; o que estamos fazendo, ao contrário, é usar apenas dados passados para caracterizar uma relação entre a frequência que um termo aparece em dois semestres diferentes (que podem ser considerados variáveis aleatórias independentes): classificando se é uma relação anômala ou não.

De acordo com a distribuição de Poisson, a probabilidade de um termo (seja simples ou composto) aparecer *k* vezes (frequência observada em um tempo *t*, no nosso caso, um semestre), é dada por[[1]](#footnote-1):

Onde:

* pmf é a função massa de probabilidade;
* é a base do logaritmo natural (e = 2, 71828...);
* *λ* é o parâmetro do modelo que indica o número esperado de ocorrências no intervalo de tempo *t*.
* *k* é o número de ocorrências em um intervalo de tempo *t*.

Diferentes escolhas do parâmetro *λ* levam a diferentes distribuições de Poisson:



Figura 1 - PMFs e CDFs para diferentes valores de λ.

A função distribuição acumulada de Poisson traz o somatório de todas as realizações da variável aleatória X até um número k de ocorrências:

Repare que a distribuição de Poisson não é simétrica. Por exemplo, para *λ*= 20, o intervalo entre k=13 e k = 28 representa 90% da probabilidade e pmf(k=13, *λ*=20)1,06% , pmf(k=28, *λ*=20)0,83%, cdf(k=13, *λ*=20)6,6% e cdf(k=28, *λ*=20)96,6% (Figura 1).

A screenshot of a cell phone

Description automatically generated

Figura 2- PMF e CDF para λ=20. A cauda direita está hachurada em vermelho e a esquerda hachurada de laranja

Sendo:

* um semestre específico.
* a frequência normalizada do termo no semestre *t*.

A forma mais intuitiva de aplicar esse modelo para o nosso problema seria, então, fazer:

* *λ = f*(termo, t-1)
* *k* = *f*(termo, t)

ou seja, assumir que o futuro será como o passado. Os termos que estão nas caudas da distribuição são anômalos. No caso intuitivo, nos interessamos pelas anomalias da cauda direita. Para obter os termos de interesse, portanto, ordenamos decrescentemente os termos em ordem de   
cdf(k, *λ* e selecionamos os 1000 primeiros termos.

Entretanto, tal aplicação intuitiva tem um problema: *f*(termo, t-1) é zero sempre que o termo aparece pela primeira vez no semestre *t* e tanto pmf(*λ*=0), quanto cdf(*λ*=0) são indefinidas. Qualquer tentativa *ad hoc* de adicionar um valor arbitrariamente pequeno a *λ* pode modificar o resultado de maneira imprevisível.

Uma solução para isso, é aplicar o modelo de uma forma não tão intuitiva, porém onde tal problema do termo com frequência nula não existe.

Aproveitando que a ocorrência de um termo em diferentes semestres podem são independentes, ao invés de calcular a probabilidade de *k* = *f*(termo, t), usando como parâmetro *λ = f*(termo, t-1), fazemos o contrário:

* *λ = f*(termo, t)
* *k* = *f*(termo, t-1)

ou seja, esperar que o passado tenha sido igual ao que sabemos do presente. Nesse caso, estamos interessados pelas anomalias da cauda esquerda. Para isso basta mudar o sentido da ordenação no método descrito para o caso intuitivo para selecionar os 1000 termos menos prováveis.

1. Sobre a distribuição de Poisson ver, por exemplo: <https://www.sciencedirect.com/topics/mathematics/poisson-distribution>. [↑](#footnote-ref-1)