Nome: Rafael Claro Ito (R.A.: 118430)
Resumo do artigo: **The Curious Case of Neural Text Degeneration**

Contexto:
O artigo explora diferentes métodos de decodificação usado para geração de textos e propõe uma nova abordagem de decodificação estocástica denominada *Nucleus Sampling*, cuja principal ideia é usar o formato da distribuição de probabilidade para se determinar o conjunto de *tokens* a serem amostrados. Com isso, aumenta-se a imunidade a degeneração, isto é, produção de textos incoerentes (palavras ilustradas em vermelho nos exemplos) e/ou ficar preso em loops repetitivos (palavras ilustradas em azul).

Geração de texto:
Para a tarefa de geração de textos são consideradas duas tarefas distintas: **geração direcionada**, onde a saída é uma transformação da entrada (ex: tradução e sumarização) e **geração *open-ended***, que envolve continuação de textos condicionados a um contexto de entrada. Todo o trabalho do artigo é baseado nesta última tarefa.

Comparação com outros métodos:
O *Nucleus Sampling* é uma abordagem parecida com a *top-k*, se diferenciando pela forma de truncamento da distribuição. Ao invés de se fixar um número fixo *k* de palavras mais prováveis para se fazer a amostragem, escolhe-se uma porcentagem *p*. Assim, a amostragem será feita em cima das palavras mais prováveis (*top-p*), tal que a soma de suas probabilidades sejam maior ou igual a *p*. Na prática esse truncamento coloca uma probabilidade zero nas palavras na cauda da distribuição, que embora individualmente apresentem baixa probabilidade, juntas acabam muitas vezes sendo significantes devido ao tamanho do vocabulário.

Uma dificuldade do método ***top-k*** é a escolha do *k* adequado. Isso acontece pois em um contexto onde a distribuição de probabilidade da próxima palavra é *flat*, um valor de *k* mais alto é adequado. Já quando a distribuição de probabilidade do próximo *token* está em poucas palavras, apenas estas devem ser usadas para a amostragem, portanto um *k* baixo é mais adequado. Ao usar o *top-p* esse problema é evitado, pois o tamanho do set a ser amostrado é dinamicamente ajustado baseado na forma da distribuição a cada passo.

Os outros métodos de decodificação comparados foram: ***beam search*** (com diversos valores de *beam width*) e ***pure sampling*** (com ou sem temperatura). Essa abordagem de *sampling* com temperatura divide o valor dos logitos por um valor entre [0,1) antes de aplicar a *softmax*. Isso ajuda a resolver o problema do *top-k*, pois diminui a massa de probabilidade na região da cauda da distribuição, mas em contrapartida diminui a diversidade.

Perplexidade:
O artigo compara a perplexidade das diferentes abordagens em relação ao texto humano enfatizando que embora uma perplexidade baixa signifique que a próxima palavra está sendo predita com maior probabilidade, isso também acaba levando a uma menor diversidade e levando a repetições. As abordagens *beam search* e *top-k* com temperatura apresentam perplexidade baixa. *Sampling*, *top-k* e *nucleus* podem podem ser calibrados para uma perplexidade humana, entretanto os dois primeiros podem ter problemas de coerência, devido aos parâmetros altos.

Resultados:
Para comparar e avaliar a performance dos métodos de decodificação de textos, foram consideradas três vertentes: avaliação probabilística (calculando a **perplexidade**), distruibuição estatística (analisando o **coeficiente Zipf** para medir uso do vocabulário, ***self-BLEU*** como métrica de diversidade, e taxa de **repetição** de sequências), e avaliação humana (computando métrica **HUSE**, como forma de medir coerência).

Os resultados podem ser vistos na tabela 1 do artigo, considerando que o objetivo não é obter a melhor pontuação na categoria, mas sim assemelhar-se ao máximo possível das pontuações de um texto humano. Por fim, os autores concluem que o *nucleus sampling* é eficaz em capturar a região de confiança de modelos de linguagem e citam que no futuro pretendem caracterizar essa região dinamicamente e incluir funções semânticas no processo de decodificação.