Planejamento Estratégico do Projeto de Webscrapping e Análise de Sentimento

Este documento apresenta o planejamento estratégico para a coleta de dados, tratamento e análise de sentimento referentes a temas como contabilidade, contabilidade digital, contabilidade online, e dúvidas sobre impostos para diferentes profissões.

---

1. Objetivos do Projeto

- **Coleta de Dados:**
  Implementar uma rotina de webscrapping para extrair informações e conteúdos (textos, comentários, posts, etc.) relacionados aos termos:
  - `contabilidade`
  - `contabilidade digital`
  - `contabilidade online`
  - `imposto para representante comercial`
  - `imposto para infoprodutor`
  - `imposto para revendedor de veículos`
  - `contabilizei`
  - `contador online`
- **Normalização dos Termos:**
  Para os itens que contenham a palavra "imposto", realizar a substituição para o padrão:
  "como declarar imposto de [nome da profissão]"
  (por exemplo, transformar "imposto para representante comercial" em "como declarar imposto de representante comercial").
- **Análise de Sentimento:**
  Após o tratamento dos dados, aplicar técnicas de processamento de linguagem natural para identificar sentimentos e dúvidas dos usuários.
- **Geração de Insights:**
  Desenvolver dashboards e relatórios para compreender as principais dúvidas e dores dos usuários em relação à contabilidade e tributação.

---

2. Etapas do Projeto

Etapa 1 – Definição do Escopo e Planejamento Inicial

- **Descrição:**
  Delimitar quais informações serão coletadas, definir as fontes priorizadas, objetivos de negócio e métricas de sucesso.

- **Atividades:**
  - Especificar a lista de termos de busca.
  - Definir as profissões para normalizar o termo "imposto".
  - Levantar requisitos técnicos e legais (respeito ao `robots.txt`, APIs, etc.).

Etapa 2 – Identificação e Seleção dos Canais de Busca

- **Descrição:**
  Mapear e selecionar os canais (websites, motores de busca, redes sociais, fóruns e blogs especializados) que serão fontes dos dados.
- **Atividades:**
  - Elaborar uma lista preliminar de canais.
  - Consultar a tabela comparativa abaixo para selecionar as melhores fontes, considerando vantagens e desafios.

Etapa 3 – Desenvolvimento da Rotina de Webscrapping

- **Descrição:**
  Projetar e implementar scripts para extrair dados dos canais escolhidos.
- **Atividades:**
  - Escolher ferramentas e bibliotecas (ex.: **Scrapy**, **BeautifulSoup**, **Selenium**).
  - Desenvolver crawlers ou scripts que acessem os conteúdos de forma ética, respeitando as políticas dos sites.
  - Implementar mecanismos para contornar bloqueios (CAPTCHAs, limites de requisição).

Etapa 4 – Armazenamento e Tratamento dos Dados (ETL)

- **Descrição:**
  Estruturar, armazenar e limpar os dados coletados, preparando-os para análises.
- **Atividades:**
  - **Extração:** Coletar os dados brutos.
  - **Transformação:**
    - Limpar e padronizar os textos (remoção de HTML, stopwords, normalização de acentuação).
    - Substituir os termos "imposto" para o formato "como declarar imposto de [profissão]".
    - Unificar formatos e corrigir inconsistências.
  - **Carga:** Armazenar os dados em um banco de dados ou arquivos estruturados (CSV, JSON, etc.).

Etapa 5 – Análise Exploratória dos Dados

- **Descrição:**
  Realizar uma análise inicial para compreender a estrutura dos dados, identificar padrões e outliers.
- **Atividades:**
  - Utilizar ferramentas de visualização (ex.: **matplotlib**, **seaborn**, **Tableau**).
  - Gerar métricas descritivas (frequência de palavras, volume de postagens, etc.).

Etapa 6 – Análise de Sentimento

- **Descrição:**
  Aplicar modelos de processamento de linguagem natural para identificar sentimentos e nuances dos textos.
- **Atividades:**
  - Selecionar bibliotecas e modelos adequados para o português (ex.: **VADER adaptado**, **TextBlobPT**, **BERTimbau**).
  - Pré-processar os textos (tokenização, remoção de ruído, etc.).
  - Treinar ou ajustar o modelo, se necessário.
  - Classificar os textos quanto aos sentimentos (positivo, negativo, neutro) e identificar principais temas.

Etapa 7 – Visualização e Relatórios

- **Descrição:**
  Criar dashboards e relatórios para visualizar os insights e facilitar a tomada de decisão.
- **Atividades:**
  - Construir visualizações interativas com ferramentas como **Power BI**, **Tableau**, ou bibliotecas Python (**Plotly**, **Dash**).
  - Elaborar um relatório detalhado com conclusões, tendências e recomendações.

Etapa 8 – Validação, Documentação e Monitoramento

- **Descrição:**
  Validar os resultados com especialistas e documentar o processo para futuras melhorias.
- **Atividades:**
  - Solicitar feedback de profissionais da área.
  - Documentar o fluxo completo (código, metodologia e desafios).
  - Estabelecer uma rotina de monitoramento e atualização dos dados e modelos.

---

3. Tabela Comparativa dos Canais de Busca para Web-scrapping

Canal	Descrição	Vantagens	Desvantagens	Exemplo de Uso
Google Trends	Plataforma que apresenta dados agregados de volume de buscas e tendências.	- Dados de tendências e sazonalidade.- Gratuito e de fácil acesso.	- Dados agregados (não traz textos completos).- Não permite extração direta para análise de sentimento.	Validar a popularidade dos termos e direcionar o foco do scraping.
Google Search (SERP)	Motor de busca que retorna resultados de diversos websites, incluindo artigos, blogs e FAQs.	- Ampla variedade de fontes e conteúdo diversificado.- Atualizações constantes.	- Possíveis bloqueios (CAPTCHAs) e restrições de scraping.- Questões legais com termos de uso.	Coletar artigos e postagens sobre contabilidade e tributação.
YouTube	Plataforma de vídeos com descrições, transcrições (quando disponíveis) e comentários dos usuários.	- Conteúdo multimídia com opiniões e dúvidas em comentários.- Grande volume de informações qualitativas.	- Extração complexa (necessidade de processar transcrições e comentários).- Formatos variados.	Extrair comentários e descrições de vídeos que abordem "como declarar imposto de [profissão]".
Redes Sociais (Twitter, Facebook)	Plataformas onde os usuários expressam opiniões e dúvidas de forma espontânea.	- Dados em tempo real e opiniões autênticas.- Possibilidade de identificar tendências emergentes.	- APIs restritivas e acesso limitado.- Questões de privacidade e dados ruidosos.	Analisar posts e tweets sobre dúvidas em contabilidade e impostos.

Canal	Descrição	Vantagens	Desvantagens	Exemplo de Uso
Fóruns e Comunidades (Reddit, Quora)	Plataformas focadas em discussões e trocas de conhecimento entre usuários.	- Conteúdo com perguntas e respostas detalhadas.- Discussões mais aprofundadas e qualitativas.	- Menor volume de dados comparado às redes sociais.- Variedade na qualidade e uso de jargões informais.	Coletar perguntas e debates sobre "como declarar imposto de [profissão]".
Blogs Especializados	Sites e blogs focados em contabilidade e tributação.	- Conteúdo especializado e aprofundado.- Alta relevância para o nicho.	- Layouts variados que exigem adaptabilidade dos scrapers.- Menor frequência de atualização.	Extrair artigos e análises especializadas sobre contabilidade digital e declaração de impostos.

---

4. Considerações Finais

- **Aspectos Legais e Éticos:**
  Respeitar as políticas de cada website (robots.txt, termos de uso e privacidade) e utilizar APIs oficiais sempre que possível.
- **Ferramentas e Tecnologias Sugeridas:**
  - **Webscraping:** Python (Scrapy, BeautifulSoup, Selenium)
  - **Armazenamento e ETL:** Pandas, bancos de dados relacionais ou NoSQL
  - **Análise de Sentimento:** NLTK, spaCy, VADER, TextBlobPT ou modelos pré-treinados para português (ex.: BERTimbau)
  - **Visualização:** Matplotlib, Seaborn, Plotly, Tableau ou Power BI
- **Iteratividade:**
  O projeto deve ser iterativo, com refinamentos contínuos na coleta, tratamento e análise dos dados conforme novas fontes e necessidades forem identificadas.

FIM DO PLANEJAMENTO ESTRATÉGICO