**CRITERIOS DE ACEPTACIÓN**

**3.1**: Recopilación de datos de texto para entrenar el modelo

* Inclusión de Habilidades Blandas: El conjunto de datos deben contener ejemplos que aborden específicamente habilidades blandas relevantes para el/los puestos IT, como por ejemplo la comunicación efectiva, trabajo en equipo, resolución de problemas, etc.
* Contextualización de Habilidades: Los datos con los que se trabajen de proporcionar contextos claros que permitan al modelo entender cómo se manifiestan las habilidades blandas en el ámbito de las ofertas laborales.
* Diversidad de Contextos Laborales: Incluir ejemplos de diversas ofertas laborales para garantizar que el modelo pueda generalizar y reconocer las habilidades blandas en diferentes contextos.
* Tamaño del Conjunto de Datos: El conjunto de datos debe ser lo suficientemente grande y variado para representar adecuadamente la diversidad de habilidades blandas que se encuentran en las ofertas laborales de IT.

Para desarrollar la tarea 3.1 de recopilación de datos de texto para entrenar el modelo, seguiría los siguientes pasos:

1. Identificar fuentes de datos de ofertas laborales IT que contengan descripciones detalladas de las habilidades y requerimientos. Algunas fuentes potenciales son sitios web de empleos, scrapped job postings, descripciones de puestos de empresas, etc.
2. Utilizar técnicas de web scraping o extracción de datos para obtener grandes volúmenes de descripciones de trabajo raw text de estas fuentes.
3. Preprocesar los datos de texto para limpiarlos - esto incluye remover HTML, eliminar duplicados, corregir errores tipográficos, etc.
4. Identificar y etiquetar secciones de los textos que se relacionen a habilidades blandas. Esto se puede hacer manualmente al inicio y luego utilizar técnicas como rules-based labeling para escalar.
5. Asegurarse que los datos de texto incluyan ejemplos específicos que cubran las habilidades blandas relevantes como comunicación, trabajo en equipo, resolución de problemas, etc.

6. Revisar los ejemplos para garantizar que proporcionen contextos claros para cada habilidad - por ejemplo, comunicación efectiva al interactuar con clientes.

7. Incluir una diversa variedad de ofertas laborales IT - backend, frontend, testing, seguridad, etc. Esto permite que el modelo generalice mejor.

8. Determinar un tamaño de conjunto de datos inicial de al menos 50,000 etiquetas o preferiblemente över 100,000 para garantizar una buena cobertura.

9. Refinar el proceso de recolección, etiquetado y procesamiento de forma iterativa para mejorar la calidad, cobertura y tamaño del conjunto de datos de entrenamiento.

Estaría iterando y ampliando el conjunto de datos hasta cumplir con los criterios de aceptación relacionados al contenido, contexto, diversidad y tamaño.

**PLAN DE TRABAJO 18/12/2023**

Aquí está una propuesta de planificación de tareas para recolectar y preprocesar los datos de texto para entrenar el modelo de habilidades blandas en 8 horas:

1. Identificar fuentes de datos (30 mins)

- Sitios web de empleos como LinkedIn, Indeed, Glassdoor

- Sitios con datos abiertos de descripciones de trabajo

- Scrapear ofertas laborales IT específicas

2. Web scraping inicial (60 mins)

- Definir queries y keywords para isolatar ofertas IT relevantes

- Crear script Python con BeautifulSoup/Selenium para extraer datos raw

- Guardar datos como JSON para facilitar preprocesamiento

```python

# Ejemplo con BeautifulSoup

import requests

from bs4 import BeautifulSoup

import json

url = "www.sitio-empleos.com"

page = requests.get(url)

soup = BeautifulSoup(page.content, 'html.parser')

results = []

for job in soup.find\_all('div', class\_='job-listing'):

title = job.find('h3').text

desc = job.find('p').text

job\_data = {'title': title, 'desc': desc}

results.append(job\_data)

with open('jobs.json', 'w') as f:

json.dump(results, f)

```

3. Preprocesamiento inicial (90 mins)

- Limpieza de datos (remover HTML, fix encoding)

- Identificar secciones relevantes (skills, requerimientos)

- Extraer habilidades blandas en contexto

- Guardar datos procesados como CSV

```python

import pandas as pd

import nltk

from nltk.corpus import stopwords

#Cargar datos raw

df = pd.read\_json('jobs.json')

# Preprocesamiento

df['clean\_desc'] = df['desc'].apply(clean\_text)

# Extraer secciones de habilidades

skills = df['clean\_desc'].apply(extract\_skills)

# Remover stopwords,limpiar

skills = [remove\_stopwords(text) for text in skills]

skills = [clean\_text(text) for text in skills]

# Guardar skills como CSV

df\_skills = pd.DataFrame(skills, columns=['skills'])

df\_skills.to\_csv('skills.csv', index=False)

```

4. Revisión y etiquetado manual (120 mins)

- Revisar muestra aleatoria para quality assurance

- Identificar y etiquetar habilidades blandas en contexto

- Guardar datos etiquetados como CSV

```python

import random

# Cargar datos limpios

df\_skills = pd.read\_csv('skills.csv')

# Muestra aleatoria para etiquetado

sample = df\_skills.sample(100)

# Etiquetado manual de habilidades blandas

labeled = []

for text in sample['skills']:

labels = # etiquetas...

labeled.append((text, labels))

# Guardar muestra etiquetada

df\_labels = pd.DataFrame(labeled, columns=['text','labels'])

df\_labels.to\_csv('labeled.csv', index=False)

```

5. Automatizar más scraping y preprocesamiento (60 mins)

- Agregar más queries y fuentes de datos

- Configurar scripts para correr periodicamente

- Revisar samples manualmente

6. Iterar sobre calidad de datos y modelos (60 mins)

- Ajustar proceso de recolección de datos según necesidades

- Entrenar modelos iniciales para validar utilidad de datos

- Recolectar realimentación para mejora continua

Esto permitiría en 8 horas tener un MVP de pipeline de datos listo y resultados iniciales para seguir construyendo iterativamente.

Tomar postulaciones desde LinkedIn como formato en texto plano para diferentes perfiles IT. Emplear esto como entrada del modelo para que el mismo pueda extraer las palabras de interés sobre soft\_skills vinculadas al perfil específico, para luego buscar los candidatos que mejor matchean con ese perfil.

**3.2**: Desarrollar el modelo de NLP para identificar habilidades blandas:

* Entrenamiento Exitoso: El modelo debe entrenarse con el conjunto de datos recopilados, demostrando la capacidad de identificar y comprender habilidades blandas en las ofertas laborales de IT.
* **Precisión y Recuperación:** El modelo debe lograr una precisión y recuperación aceptables al identificar habilidades blandas, minimizando tanto los **falsos positivos** como los **falsos negativos**.
* **Interpretación de Contexto:** El modelo debe ser capaz de interpretar el contexto específico de las ofertas laborales en IT para **comprender** adecuadamente **cómo** se manifiestan las habilidades blandas en diferentes situaciones.
* Manejo de Sinónimos y Variaciones Lingüísticas: El modelo debe ser robusto ante sinónimos y variaciones lingüísticas de las habilidades blandas, reconociéndolas independientemente de la forma en que se expresen.