HTML Clones

Pre-Intro:

Pentru a ma calibra corect in raport cu cerinta, am pornit de la presupunerea ca nu se doreste identificarea exclusiva a paginilor complet identice, ci a celor care sunt *vizual similare*, chiar daca pot exista mici diferente nesemnificative, diferente care, in majoritatea cazurilor, nici nu ar fi sesizate de ochiul uman la prima vedere.

Intro:

Salutare, Domnilor!
Numele meu este Cosmin-Sergiu Milica, sunt absolvent al Facultatii de Automatica si Calculatoare si am acumulat o mica experienta profesionala in cadrul Bitdefender. Am ales acest task pentru ca mi s-a parut interesant si provocator, oferind multiple directii de rezolvare. Am decis sa abordez problema dintr-un unghi putin mai atipic, sa fie o solutie diferita, nu neaparat prima varianta care iti vine in minte sau pe care o sugereaza ChatGPT.

Am pornit de la cunostintele generale acumulate si de la inclinatia mea catre zona de cybersecurity. M-am gandit: daca vreau sa compar doua lucruri, ce metoda imi ofera o evaluare rapida si eficienta? Raspunsul meu automat a fost: hash-uri.

Apoi am analizat ce este de fapt un DOM HTML=o structura de tip arbore. Pentru a putea compara eficient astfel de structuri, aveam nevoie de o reprezentare eficienta, care sa-mi permita parcurgere si comparare recursive. Asa am ajuns sa folosesc o abordare inspirata din Merkle Trees.

Un Merkle Tree este un arbore de hash-uri construit de jos in sus: fiecare nod frunza reprezinta un hash al unei componente, iar nodurile intermediare contin hash-uri compuse din hash-urile copiilor. Avantajul acestui model este ca poti identifica rapid, printr-o binary search, care parte din structura a fost modificata.

Pentru hashing am ales sa folosesc **fuzzy hashing** , in locul unor functii clasice precum SHA256. Motivul este simplu: functiile de hash standard sunt prea stricte, orice diferenta minuscula rezulta intr-un hash complet diferit. In schimb, fuzzy hashing imi permite sa detectez similaritati partiale intre doua structuri, ceea ce este esential in contextul paginilor HTML, unde diferentele minore nu ar trebui sa duca la excluderea dintr-un grup similar.

Solitie:

1. Normalizarea paginilor HTML

Prima intrebare pe care mi-am pus-o a fost: Ce insemana cu adevarat stuctura vizuala a unei pagini HTML?. Am constat ca nu pot compara codul sursa brut, deoarece doua pagini pot arata identic in browser, dar sa fie scrise complet diferit. Asa ca am inceput un proces de normaliza, eliminarea a noise-ului, din paginile HTML.
Am eliminat toate tag-urile care nu contribuie la randarea efectiva: `script`, `meta`, `link`,

`noscript`, etc. Acestea ar fi incarcat inutil arborele DOM si ar fi introdus diferente irelevante.
Pentru a uniformiza stilurile, am extras toate regulile CSS inline si le-am aplicat direct pe elementele HTML. Astfel, fiecare element pastreaza doar stilurile care il afecteaza vizual, indiferent de sursa acestora.
Am pastrat doar atributele vizuale esentiale: `style` si `class`. Alte atribute precum `id`, `href`, `data-*` au fost eliminate deoarece nu contribuie la layout-ul vizual.
Textul intern a fost curatat de spatii redundante si formate neregulate, retinand doar prezenta textului ca semnal vizual.

Acest pas standardizeaza reprezentarea vizuala a paginilor HTML, reducand variatiile inutile si pastrand doar esenta care conteaza pentru comparare.

2. Merkle Tree

Pentru a compara structura normalizata a unei pagini HTML, am implementat un arbore inspirat din **Merkle Tree**. Fiecare element HTML este transformat intr-un nod, care contine atat informatii vizuale, cat si hash-uri care ajuta la comparare eficienta.

Structura unui nod:

- *tag*: tag-ul HTML
- *attrs:* atributele vizuale
- *text:* un fragment scurt din textul continut de nod (maxim 15 caractere), doar pentru a semnala prezenta sa vizuala, nu pentru continut semantic
- *children:* lista cu nodurile copil
- *fingerprint:* o semnatura textuala ce unifica tag, class, style si text, este esenta vizuala a nodului
- *hash:* fuzzy hash-ul (ssdeep) al fingerprint-ului pentru nodurile frunza sau al concatenarii hash-urilor copiilor pentru nodurile interne

Exemplu de fingerprint:

```HTML
<div|id=|class=container
main|style=color:red;font-size:14px|text=Hello world>
```

Calculul hash-ului:

- Pentru frunze: `ssdeep.hash(fingerprint)`
- Pentru noduri interne: `ssdeep.hash(concat_hashuri_copii)`

Astfel, orice modificare la nivel de structura sau continut se propaga in hash-urile superioare din arbore, facand posibila detectarea eficienta a diferentelor, fara a compara intregul DOM manual. Am ales `ssdeep` pentru ca imi permite sa obtin un scor de similaritate intre doua fingerprint-uri. Desi poate avea coliziuni, avantajul compararii partiale este esential pentru detectarea similaritatilor intre structuri mari

Logica de construire a arborelui:

- Daca nodul este doar text brut (adica nu are un tag asociat), extragem continutul textului, il curatam de spatii si il convertim intr-un nod special de tip `__text__`. Daca textul e gol, il ignoram complet
- Daca nodul este un element HTML valid, extragem atributele vizuale (`class`, `style`, etc.) si continuam recursiv pe toti copiii sai din DOM, construind lista de noduri copil
- Constructia se face recursiv bottom-up, ceea ce inseamna ca intai se genereaza nodurile frunza, apoi se compun nodurile parinte, care contin hash-uri bazate pe hash-urile copiilor. La final, obtinem un arbore complet care reprezinta toata pagina HTML intr-o forma compacta si comparabila

3. Clusterizare

Dupa ce am obtinut o reprezentare structurata pentru fiecare pagina HTML sub forma de Merkle Tree, obiectivul final este sa identificam grupuri de pagini care sunt vizual similare. Pentru asta avem doua etape majore:

1. Compararea arborilor (Merkle Tree Matching)
   a. Pasul 1: Comparam hash-urile fiecarui nod
      Fiecare nod are un fuzzy hash (`ssdeep`), construit dintr-un fingerprint vizual (`tag`, `style`, `class`, `text`). Compararea dintre hash-urile a doua noduri returneaza un scor intre `0 si 100`.
      Daca scorul este >= threshold (default 80), consideram ca nodurile sunt similare, nu continuam recursiv pe copii.
   b. Pasul 2: Comparare de backup – fingerprint similarity
      Daca hash-ul nu este suficient de asemanator, intram mai in profunzime. Compar fingerprint-urile celor doua noduri folosind `SequenceMatcher` din `difflib`, care returneaza un scor de similaritate procentual intre string-uri. Daca scorul este >= threshold + 10 (adica 90), il consider un match vizual acceptabil, chiar daca hash-ul a picat.
   c. Pasul 3: Comparare recursiva a copiilor
      Daca nodurile nu se potrivesc dupa pasul 1 si 2, si daca au copii, algoritmul continua recursiv si compara fiecare pereche de copii de pe aceeasi pozitie (`i`), in paralel.
      - Daca gaseste un `missing node` (exista la A dar lipseste la B sau invers), adauga o diferenta.
      - Daca un nod frunza este diferit vizual, adauga o diferenta cu explicatie (`leaf mismatch`).

2. Clusterizarea paginilor
Acum ca stim cum sa comparam doua pagini, vine intrebarea: cum le grupam?

- `max_allowed_diffs` este un parametru cheie (default 3) care spune cat de multe diferente sunt tolerate intre doua pagini.
- Paginile care au suficiente asemanari intra in acelasi grup, iar cele care difera prea mult raman singure sau formeaza alte grupuri.

Acest model a fost gandit sa functioneze in felul urmator:

- Sa fie tolerant la diferente mici de stil si layout
- Sa functioneze bine chiar si cand codul este scris diferit, dar randarea vizuala este asemanatoare
- Scorurile fuzzy si compararea structurala sa mearga mana in mana: unul pentru precizie rapida, celalat pentur fallback sigur

Parametrii principali ai algoritmului, `threshold` pentru fuzzy hashing si `max_allowed_diffs` pentru toleranta la diferente, au fost stabiliti in urma testarii pe un set-urile de date primite.

Alte incercari:

Am testat si o varianta alternativa care implica realizarea de screenshot-uri ale paginilor HTML, urmate de compararea acestora folosind perceptual hashing. Desi ideea aducea rezultatele, am renuntat din doua motive principale:

1. Timpul de procesare era ridicat, pentru ca fiecare pagina trebuia randata complet folosind un headless browser.
2. Solutia era destul de costisitoare in termeni de memorie si stocare, deoarece era necesar sa salvez screenshot-urile pentru a le putea compara ulterior.

Rezultate:

Tier1

Group 1: ['alphamaterialsinc.com.html', 'pvcgs.org.html', 'jandptrucking.com.html', 'americanairless.com.html', 'keepmybooks.services.html', 'keepmybooks.pro.html', 'doughansonconstruction.com.html', 'fidexor.com.html', 'citizensagainstsextrafficking.org.html', 'ordfld.com.html', 'nounsbverbn.com.html', 'moneyweedwives.show.html', 'moneyweedwives.com.html', 'grantiah.com.html', 'angelvisiontravel.com.html', 'pyramidelectric.us.html', 'rovics.com.html', 'crazyadsclimber.com.html', 'brakeditorial.com.html', 'harotzu.com.html']

Group 2: ['alisupermercato.eu.html', 'asahibeerusa.com.html', 'asiafundspace.com.html', 'arcadeeurope.com.html', 'approvedfast.com.html', 'alileime.org.html', 'almacom-gmbh.eu.html', 'alinahoivatiimi.com.html', 'apimco.link.html', 'ahamconsumerconnections.org.html', 'ai-center.online.html', 'alisupermercati.com.html', 'aigner-haag.at.html', 'ahbynmkkmnfu.shop.html', 'alimarkets.it.html', 'aotvqsuprqnb.shop.html', 'angangintl.com.html', 'aerex.eu.html', 'app-go88s.biz.html', 'amdac-carmichael.com.html', 'aevesdk3.com.html', 'afro-pari.com.html', 'alessiofalcone.it.html', 'alwin.ltd.uk.html', 'alinahoivatiimi.net.html', 'annabeodog.xyz.html', 'arbetslivsmuseer.se.html', 'alimarkets.com.html',

'almighty-jezuz.com.html', 'akashinime.guru.html', 'arabianchemicalterminals.com.html', 'apps-foundry.com.html', 'audreysweets.xyz.html', 'aliper.it.html', 'appleclub.tech.html', 'aemails.org.html', 'arttoy.cc.html', 'apco911.com.html', 'alhasanfoundation.in.html', 'asd.net.html', 'argonfinancial.com.html', 'aliper.com.html']

Group 3: ['concoursparcscanada.ca.html', 'datewithdice.com.html', 'amyqnliycusz.shop.html', 'badlandsconcerts.com.html', 'globewayimmigration.com.html', 'columbiahouse.ca.html', 'membranereactor.com.html', 'dudecheck.com.html', 'wifipresspad.com.html', 'badlandslightfest.com.html', 'fmdistilled.com.html', 'templarsnotary.com.html', 'nobullheating.com.html', 'rootsbluesbarbecue.com.html', 'couplesdash.com.html', 'soultosolesoundspa.com.html', 'ilovestubbs.com.html']

Group 4: ['sodearif.com.html', 'golf-saint-cyprien.com.html', 'championdirect.store.html']

Group 5: ['amcun9.online.html', 'amcun3.online.html']

Group 6: ['aitoka.shop.html']

Group 7: ['ashfordcenter.world.html']

Group 8: ['astroservice.top.html']

Group 9: ['amordevoltarapido.com.br.html', 'atyourlevel.online.html']

Group 10: ['paddygower.com.html', 'eonfibre.net.html', 'scalingspecialists.com.html', 'stratalaser.com.html', 'babubasics.co.uk.html', 'thisisthefuckingnews.com.html', 'babubasics.com.html']

Group 11: ['anzald.com.html']

Group 12: ['amt-avaluos.online.html']

Group 13: ['authologic.io.html']

Group 14: ['celestialkeepsakes.com.html']

Group 15: ['artfay.tv.html']

Tier2

Group 1: ['acco-semi.com.html', 'healthfly.in.html', 'bestcontentwritingservice.com.html', 'tiptopteak.com.html', 'djdrinks.co.uk.html', 'creplace.com.html', 'engineeredrss.com.html', 'nykei.com.html', 'fortunatextiles.es.html', 'mariner-energy.com.html']

Group 2: ['shopmeds.us.html']

Group 3: ['alessiodecurtis.com.html']

Group 4: ['petapilot.com.html']

Group 5: ['kroha.de.html', 'local-marketing-lab.com.html', 'ads-sedlmair.online.html', 'hhammer.de.html', 'hanshammer.de.html', 'techcom-gmbh.de.html']

Group 6: ['starkwelt.com.html']

Group 7: ['mpgsx.com.mx.html']

Group 8: ['altenheime-essen.de.html']

Tier3

Group 1: ['ampika.com.html']

Group 2: ['eastbourne.online.html', 'belledermaaesthetics.com.html', 'thefoxinnbroadwell.com.html', 'adeptohomes.com.html', 'lipolondon.com.html', 'lagustosavaldebebas.com.html', 'deltadoula.com.html', 'hycareakarui.com.html', 'renewconsultants.com.html', 'fieldtech.info.html', 'renautautomotive.com.html', 'sophrologue-paris14.com.html', '1stmortgages.london.html', 'londonviptaxi.com.html', 'okcis.info.html', 'flyerfunnelsuk.com.html', 'frankieswinebar.com.html', 'g3rcq.com.html', 'furniturehutuk.com.html', 'columbiacouncilofneighborhoods.com.html']

Group 3: ['proapremium.id.html']

Group 4: ['elitemajesty.com.html', 'parascerah.store.html', 'bersamaetawalin.site.html', 'juraganstore.shop.html', 'susuetawalinkuid.site.html', 'dhavinastore.com.html', 'afnanstore.shop.html', 'masjidrayaalfalah.id.html', 'tulangsendietawalin.site.html', 'cameliastore28.my.id.html', 'rakarta.com.html', 'etawalinherbalmilk.site.html']

Group 5: ['dvnbysarah.com.html']

Group 6: ['omerta-inc.com.html']

Group 7: ['imzcr.me.html', 'coade.icu.html']

Group 8: ['cosekarang.com.html']

Group 9: ['dianessidewalkdeli.com.html']

Tier4

Group 1: ['visittlfl.com.html', 'wowlklnd.com.html', 'novltyclub.com.html', 'shoalsslty.com.html', 'ascendohio.com.html', 'lmeloyalty.com.html', 'tlflhasit1.com.html', 'altnloyalty.com.html', 'nbesloyalty.com.html', 'zenlfs.com.html', 'tryhrbyloyalty.com.html']

Group 2: ['coronadynamics.com.html', 'your-pc-guru.com.html', 'lbgenetics.com.html', 'morriskamlay.com.html', 'ecoled.co.in.html', 'assuredrxservices.com.html', 'rmhaddock.com.html']

Group 3: ['cazinoz1-win.pp.ru.html', '1wincasinoz-vhod.org.ru.html', '1-win-cazinos-club.org.ru.html', '1win-sloty.pp.ru.html', 'vulcan-24kasinos.pp.ru.html', 'mirror-wulkan-russia.org.ru.html', '1win-official-site-casinoz.org.ru.html', 'fontan-zercalo.net.ru.html', 'fontan-mobile.net.ru.html', 'kasinos-1-win.net.ru.html', 'cazinos-official-1win.net.ru.html', 'fontan-casino.pp.ru.html']

Concluzie:
Task-ul a fost o provocare interesanta, care m-a pus in fata unei probleme reale si m-a fortat sa gandesc o solutie care sa ofere un echilibru intre precizie si eficienta in timp.
Chiar daca mai exista unele false positive (mai ales la tier 2), sunt constient de ele si intentionez sa lucrez in continuare la rafinarea algoritmului pentru a le reduce cat mai mult.

Sper ca metoda propusa a reusit sa va starneasca interesul si ca imi veti oferi un feedback constructiv, sunt aici ca sa invat, sa cresc si sa devin mai bun.
Va multumesc pentru timpul acordat si sper sa ne auzim/vedem la interviu! 😊